Baltic J. Modern Computing, Vol. 10 (2022), No. 4, 574-610
https://doi.org/10.22364/bjmc.2022.10.4.01

Systematic Review of Functional Pathways and Methods for COVID-19 Modelling

Ramunė VAIŠNORĖ[1], Gabija MAZUR[2], Violeta MIKŠTIENĖ[2], Audronė JAKAITIENĖ[1,3]

[1]Institute of Data Science and Digital Technologies, Vilnius University
[2]The Biobank of Lithuanian population and rare disorders, Institute of Biomedical Sciences, Faculty of Medicine, Vilnius University
[3]Department of Human and Medical Genetics, Institute of Biomedical Sciences, Faculty of Medicine, Vilnius University

ramune.vaisnore@mif.stud.vu.lt, gabija.mazur@mf.vu.lt, violeta.mikstiene@mf.vu.lt, audrone.jakaitiene@mf.vu.lt

Abstract. The outcome of human viral infections is highly dependent on the host features. The scale of COVID-19 spread and amount of deaths caused motivate scientists to search for ways to combat this pandemic. We have reviewed 34 scientific papers taking into account two main points of COVID-19: the biology behind the infection and the methods used to model the outcome of the disease. The findings of the studies suggest that host genetic factors impact the clinical manifestation and outcome of COVID-19. Scientists are modelling COVID-19 using various computational methods, including genome-wide, exome-wide, and phenome-wide association analyses. Machine learning and some other methods are used to model COVID-19 to obtain new insights into the pathogenesis of the disease. As for now, there is a limited number of causal studies about COVID-19 and host genetic factors.

Keywords. COVID-19, Functional Pathways, Association Studies, COVID-19 Modelling, Machine Learning, Artificial Neural Networks, Disease Prediction

1. Introduction

Recently, the world has endured the global crisis induced by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The COVID-19 pandemic has resulted in the infection of over 410 million people worldwide (on the 14th of February 2022), of whom 5.81 million have died (WHO, 2022). Although showing a declining trend, the infection remains dangerous, claiming thousands of lives every day and tending to become seasonal in the future.

COVID-19 is a very complex disorder affecting multiple organs: the respiratory system (Zhao et al., 2020), the circulatory system (Teuwen et al., 2020), the heart, the brain and the central nervous system (Coony, 2020), the renal system (Argenziano et al., 2020), and the gastrointestinal system (Ding et al., 2020). The manifestation ranges from mild to severe, and these outcomes are highly dependent on the characteristics of an individual (Kenney et al., 2017, Tahamtan et al., 2020). Demographic characteristics (age, sex, ethnicity), comorbidities, some clinical symptoms, specific laboratory test findings, and diet/lifestyle predetermine a patient's risk of developing the more severe course of the disease. Additionally, there is considerable evidence that genetic features of human genomes have a clear association with COVID-19 severity (Zhang et al., 2020; Wang D. et al., 2020; COVID-19 Host Genetics Initiative 2021).

Unravelling the host genomic factors associated with the clinical characteristics of COVID-19 can lead to a better understanding of COVID-19 development and improve disease management. In this overview, we provide the synthesis of current scientific knowledge in fundamental pathophysiology underlying COVID-19 and the computational methods/software used in the disease modelling, thus applying the system biology approach and enabling versatile and comprehensive perception of the topic complexity. To our knowledge, it is the first paper which brings insights from the field of medicine/biology together with computational modelling.

2. Literature search strategy

We present the review of overall 34 scientific articles selected by two independent scientists: one seeking to explain the biological part of COVID-19 infection, while the other reviewing the computational part of the COVID-19 outcome and severity modelling. Seven of the reviewed studies did not separate the genetic variants or the associations found were not significant, and therefore these studies are not mentioned in the part of the COVID-19 biology analysis of this review. In this review, we have used the scientific articles published before the 14th of February 2022.

For the biological part, scientist 1 searched for the articles using the following keywords: "*covid-19"*, "*sars-cov-2*", "*coronavirus*", "*genetic variation*", "*gene*", "*genome-wide association study*", "*polymorphisms*", "*single nucleotide*", "*genetic association", "genetic susceptibility", "genotype", "human host", "genotype".*

For the computational part, scientist 2 searched for the articles using the following keywords: "*covid-19 outcome modelling*", "*covid-19 severity modelling*", "*machine learning for covid-19 modelling*", "*covid-19 prediction using genomic data*". The literature search was carried out in the following databases: PubMed, medRxiv, and bioRxiv. We analyze only those articles that first described the SNP associations found.

3. The biology behind the COVID-19 infection

SARS-CoV-2 is a single-stranded (ss) RNA virus. Its genome contains 29,881 nucleotides in length and encodes 9,860 amino acids (Huang et al., 2020). The first 2/3 of the viral genome base pairs are called open reading frame sequences (ORFs). SARS-CoV-2 has two of them – ORFa and ORFb, which encode two polyproteins (pp1a and pp1ab), and other non-structural proteins (NSP). The rest of the SARS-CoV-2 genome is composed of 4 genes: S, E, N and M (visualised in Fig. 1), which encode four main structural proteins: spike (S), envelope (E), nucleocapsid (N), and membrane (M) proteins (Tavasolian et al., 2021; Kang et al., 2020).

The surface of SARS-CoV-2 is covered in a large number of S proteins. They are highly conserved and involved in receptor recognition and viral attachment to host cells (here, we focused only on a human). The N protein is a multifunctional RNA-binding protein necessary for viral RNA transcription and replication. The M protein is the most abundant structural protein and defines the shape of the viral envelope and organises the new SARS-CoV-2 assembly, interacting with all structural proteins. The E protein is the smallest of the major structural proteins. The SARS-CoV-2 genome replication cycle is associated with the assembly of new virions, effective virion transfer to new cells, and reduced stress response by the host cell (Schoeman and Fielding, 2019).

3.1. Entry and replication cycle

SARS-CoV-2 uses the S glycoprotein to promote entry into the host cell. This protein contains two functional domains: an S1 receptor-binding domain and an S2 domain that mediates the fusion of viral and host cell membranes. The S protein binds to the ACE2 receptor on the host cell, initially through the S1 receptor binding domain. The S1 domain is then shed from the viral surface, allowing the S2 domain to fuse to the host cell membrane. This process depends on the activation of the S protein by cleavage via the host protease TMPRSS2 and other proteases (coloured blue in Fig. 1) (V'kovski et al., 2021).

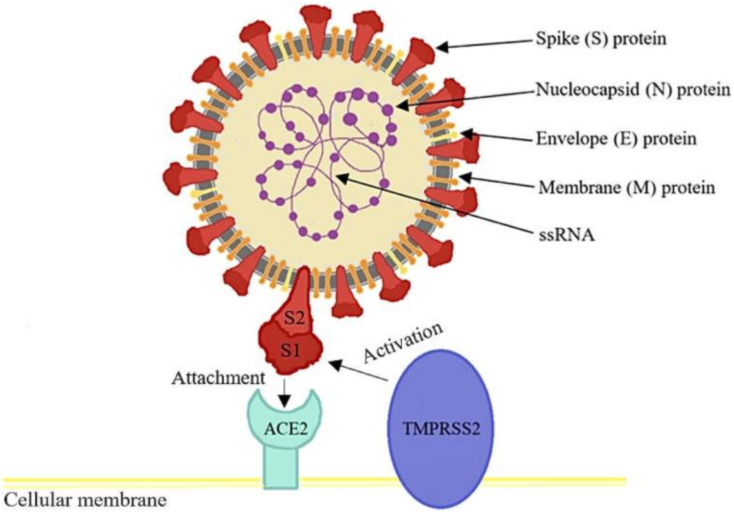

Figure 1. Schematic representation of SARS-CoV-2 architecture and spike protein (S) binding to ACE2 receptor, mediated by TMPRSS2 protease. Based on Huang et al. (2020).

Despite the short time since the beginning of the outbreak, some host genome variants have been linked with COVID-19 presentation. It has been reported that the S protein and ACE2 binding affinity are correlated with disease severity in SARS-CoV-2 infections (V'kovski et al., 2021). ACE2 gene polymorphisms that putatively increase or decrease susceptibility based on virus interactions with the S glycoprotein on the cell surface were recently described (Suryamohan et al., 2020) (Table 1). Additional locus in 3p21.31 contains several genes (*SLC6A20*, *LZTFL1*, *FYCO1*, *CXCR6*, *XCR1*, *CCR9*), and some take part in cellular biology immunity response or interact with ACE2. Variations in TMPRSS2 – a protein-encoding gene involved in SARS-CoV-2 penetration to host cell has been associated with the increased expression of TMPRSS2 on the cell surface, leading to inhibition of antiviral response (Asselta et al., 2020).

The release of the coronavirus genome into the host cell cytoplasm initiates the highly regulated onset of a complex viral gene expression program. SARS-CoV-2 has a highly conserved genomic organisation, with a large replicase, an enzyme that catalyses

RNA replication from an RNA template. The first step in the coronavirus lifecycle is the translation of the replicase from the virion genomic RNA.

Table 1. Genome variants affecting the entry of SARS-CoV-2 into the host cell.

Publication	**Gene**	**Number of variants**	**COVID-19 disease severity**	**Sample size (cases/controls)**	**Laboratory methods**
Stawiski et al., 2020	*ACE2*	9	Increase	290,000 (NA / NA)	Genotyping
		17	Decrease		
Benetti et al., 2020		3	Decrease	389 (131/258)	Whole exome sequencing
		2	Decrease		
		3	Decrease		
Guo et al., 2020		2	Increase	141,456 (NA / NA)	Whole exome and genome sequencing
		7	Decrease		
Gibson et al., 2020		5	Decrease	141,456 (NA / NA)	Whole exome and genome sequencing
		4	Increase		
Horowitz et al., 2020		1	Decrease	756,646 (52,630/704,016)	SNP genotyping assay
MacGowan and Barton, 2020		1	Increase	-	Whole exome and genome sequencing
		3	Decrease		
Hou Yuan et al., 2020	*TMPRSS2*	1	Increase	81,000 (NA / NA)	Whole exome and genome sequencing
Wulandari et al., 2021				95 cases	SNP genotyping assay
Monticelli et al., 2021		1	Decrease	1177 cases	Whole exome sequencing
Grimaudo et al., 2021	*TLL-1*	1	Increase	383 cases	SNP genotyping assay

NA – Information Not Available

The coronavirus genomic RNA encodes non-structural proteins (NSPs) critical in viral RNA synthesis and structural proteins, which are essential for virion assembly. First, the polyproteins pp1a and pp1ab are translated into functional NSPs as RNA replicase. RNA replicase is responsible for the replication of structural and non-structural protein RNA. Structural proteins S1, S2, envelope (E), and membrane (M) are translated by host ribosomes that are bound to the endoplasmic reticulum (ER) and presented on its surface as preparation for virion assembly. The nucleocapsids (N) remain in the cytoplasm and are assembled from genomic RNA. They fuse with the virion precursor, which is then transported from the host ER through the Golgi apparatus to the cell surface through small vesicles (Fehr and Perlman, 2015). These new virions are now accessible to infect another healthy cell and can also be released into the environment via respiratory droplets, potentially spreading to healthy individuals (Shah et al., 2020).

Comprehensive data summary of the studies presenting the variants interacting with the entry of SARS-CoV-2 into the host cell is provided in Supplementary Table S1.

3.2. Immune response

Once the virus gets inside the target cell, such as in the epithelial cell of the lungs, its number increases exponentially. When a certain number is reached, the host immune system recognises the changed environment and locates the virus or its surface antigenic determinants, and epitopes, inducing an immune response. Host defences come into play to block or inhibit initial infection, protect cells or eliminate virus-infected cells. The immune system comprises two main parts: innate (general) and adaptive (specialised).

Innate immune defences are initiated via pathogen recognition receptors (PRRs), and toll-like receptors (TLRs). TLRs are present on the surface of immune cells such as dendritic cells, macrophages, lymphocytes, and parenchymal cells. These receptors promote the expression of the immune cells' communication proteins called cytokines - interferons (IFN), chemokines, and others. IFNs activate natural killer (NK) cells. NK cells can kill infected cells. Chemokines may also play an essential role in innate antiviral defence by regulating macrophage, neutrophil, dendritic cells (DC), and NK responses at the site of infection (Mueller and Rouse, 2008).

The initiation of adaptive immunity is dependent on innate immunity. Innate immunity generally slows the virus, allowing the adaptive immune response to begin. Adaptive immunity leading players are lymphocytes: B cells and T cells. Adaptive immunity involves virus-specific antigen responses that are highly adapted to the specific pathogen and are firmly regulated by cross-talk between innate immune cells. Innate immune produced cytokines are drawn into lymphoid tissues, virus antigen-presenting cells (APC), and lymphocytes. APCs are immune cells that specialise in presenting a virus antigen through their major histocompatibility complex (MHC) class I and class II proteins, also known as human leukocyte antigen (HLA) (Wieczorek et al., 2017). The primary type of professional APCs is dendritic cells (DC). T and B cells are activated when they recognise foreign antigens presented by MHC proteins from APC cells. Activated B cells initiate high-affinity, antibody-producing long-lived plasma cells (mature B cells) and memory B cells. Antibodies generally function by binding to free viral particles and blocking host cell infection. Antibodies, also known as immunoglobulins (Ig), are composed of two heavy chains (H) and two light chains (L). There are five main classes of heavy chain domains. Each category defines IgM, IgG, IgA, IgD, and IgE isotypes (Schroeder and Cavacini, 2010). Memory B cells circulate throughout the body until a specific antigen is re-encountered, triggering an immune response. Activated T-cells recognise and destroy virus-infected cells (Mueller and Rouse, 2008).

While the immune system is there to protect itself, it can cause harm and requires strict regulation. The findings from studies have suggested that SARS-CoV-2 can suppress IFN signalling and impair viral clearance from infected cells. Research on SARS-CoV shows that multiple viral structural and non-structural proteins antagonise interferon responses. Antagonism occurs at various stages of the interferon signalling pathway, including preventing PRRs recognition of viral RNA. GWAS performed by Spanish researchers identified a signal located in 9q34 within the ABO blood group locus suggesting a possible association of the disease severity with blood groups (Ellinghaus et al., 2020). The meta-analysis of genetic variation implicated in excessive release of cytokines (IL-6, IL-1β, TNFα) ("cytokine storm") revealed an association of the 174C allele of the IL6 gene (and a higher level of IL-6) with the severity of pneumonia (Ulhaq and Soraya, 2020). Several pieces of research demonstrated that lacking Toll-like receptor genes led to increased viral replication and enhanced lung pathology (Ovsyannikova et al., 2020). An identified splice variant of the *OAS1* gene confers protection against COVID-19 in people of African ancestry. *OAS* genes activate viral RNA degradation and other antiviral defence mechanisms (Huffman et al., 2021) (Table 2).

Table 2. Genome variants of human immune response to SARS-CoV-2 associated with increased COVID-19 severity.

Publication	Gene	Number of variants	COVID-19 disease severity	Sample size (cases/controls)	Laboratory methods
Secolin et al., 2021	*HLA*	3	Increased	386 (NA/NA)	Whole exome sequencing
Wang D. et al., 2020		2		3,872 (82/3,790)	Next-generation sequencing
Zhang et al., 2020	*IFITM3*	1		80 (NA/NA)	*IFITM3* sequencing
Grimaudo et al., 2021	*PNPLA3*	1		383 (NA/NA)	SNP genotyping assay
Kuo et al., 2020	*APOE*	2		322,948 (622/323,570)	SNP genotyping assay
Severe Covid-19 GWAS Group, 2020	*ABO, LZTFL1*	2		3,815 (1,610/2,205)	SNP genotyping assay
Rescenko et al., 2021	*LZTFL1*	3		2,692 (475/2,217)	SNP genotyping assay
The COVID-19 Host Genetics Initiative, 2021	*SLC6A20, ABO, RPL24, PLEKHA4, LZTFL1, FOXP4, TMEM65, OAS1, KANSL1, TAC4, DPP9, RAVER1, IFNAR2*	13		2,049,562 (49,562/2,000,000)	SNP genotyping assay
Pairo-Castineira et al., 2021	*LZTFL1, CCHCR1, OAS3, DPP9, RAVER1, IFNAR2*	6		10,056 (1,676/8,380)	SNP genotyping assay
Ma et al., 2021	*SLC6A20, ABO, IFNAR2-IL10RB*	4		680,128 (3,288/676,840)	SNP genotyping assay
Shelton et al., 2021	*SLC6A20, ABO*	2		114,240 (12,972/101,268)	SNP genotyping assay
Hu et al., 2021	*DNAH /SLC39A10, CLUAP1, DES/SPEG, STXBP5, TOMM7, WSB1, PCDH15, CPQ*	23		1,096 (292/804)	Genotyping
Pairo-Castineira et al., 2021	*OAS1-3, TYK2, DPP9, IFNAR2*	4		100,000+ (2,244/100,000+)	Genotyping, whole genome sequencing
Verma et al., 2021	*ABO, RAVER1*	3		455,683 (NA / NA)	Genotyping
Fallerini et al., 2021	*TLR7*	6		156 (79/77)	Genotyping
Tanimine et al., 2021	*OAS1, IL1B*	2		230 (NA/NA)	Genotyping
Dapeng Wang et al., 2022	*EFCAB4B*	3		500,000 (10,118/489,882)	Genotyping
Maes et al., 2022	*NLRP3*	2		528 (NA/NA)	Genotyping
Carapito et al., 2021	*ADAM9*	1		72 (47/25)	Gene expression
Huffman et al., 2021	*OAS1*	1	Decreased	120,473 (1,842/118,631)	Genotyping

Recent large-scale studies have shown the role of rs1990760 (p.Ala946Thr) of the *IFIH1* gene in SARS-CoV-2 infection. Individuals carrying the T allele may be more resistant to SARS-CoV-2 infection (Maiti et al., 2020).

The findings of Kurki et al. (2021) on the Finnish cohort confirm the findings of Kuo et al. (2020) that the *APOE ε*4 allele (*APOE4*) is a risk factor for severe COVID-

19. The *APOE* immunomodulatory agent affects both innate and adaptive immune responses by stimulating macrophages and other inflammatory response immune cells, such as suppressing T cell proliferation and activating neutrophils. Moreover, Kurki and colleagues revealed the relation between *APOE4* and post-COVID mental fatigue.

More details about the studies that have found variants associated with host immune response to SARS-CoV-2 infection are listed in Supplementary Table S2.

4. Methods to model the outcome and severity of COVID-19 infection

While analysing 34 studies from the modelling perspective, we can discriminate all applied methods into three main groups: association studies, machine learning techniques, and other statistical methods. We observe that association methods were the most common, and machine learning techniques were used less frequently. Some studies have applied other statistical modelling techniques. Twenty-nine studies used a single method; in 5 publications, the authors report multiple methods (see Fig. 2.). Next, we will discuss the models of each group in more detail.

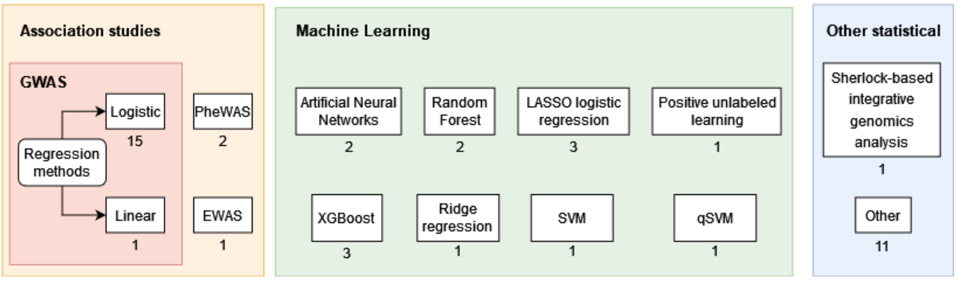

Figure 2. Main model groups for modelling the outcome of COVID-19: association studies, machine learning, and other statistical methods. The number below the title is the frequency the method was used. Some studies implemented multiple models. Therefore, the sum of the methods is greater than the studies reviewed.

4.1. Association studies

Of the 34 articles analysed, various association techniques were commonly used to model the COVID-19 disease (Table 3). Sixteen studies are case-control studies, and, respectively, 18 model the different severity of COVID-19 by only analysing the patient group. GWAS was applied in COVID-19 studies to discover genetic factors associated with the severity, mortality, risk, laboratory, and clinical characteristics of COVID-19. Scientists are studying patients of different ancestry, considering different factors such as age, sex, comorbidities, and others. In the studies, the sample size was very different from the most minor (230 patients) in Tanimine et al. (2021) to the largest (5.37 million cases and control) in the COVID-19 Host Genetics Initiative (2021) (COVID-19 HGI). (COVID-19 HGI) is a scientific collaboration seeking to spread the research findings and knowledge about the genetic basis of COVID-19 susceptibility, severity, and clinical outcomes. Currently, 115 registered studies participate in this initiative and constantly update the knowledge base with the newest findings. All studies can be summarised as having solved two different problems, that is, case-control studies want to find out what factors are associated with COVID-19 outcome, and studies that only include data from patients aim to detect what factors are associated with a different course of COVID-19 disease (see Table 3). Next, we will discuss the studies of associations found now in detail.

Genome-wide association study (GWAS)

In GWAS, scientists apply statistical tests to determine statistically significant associations between SNP alleles and phenotypes, and multiple tests are conducted simultaneously. GWAS requires three essential elements: (1) sufficiently large study samples from populations that effectively provide genetic information regarding the research question, (2) polymorphic alleles that can be inexpensively and efficiently genotyped and cover the whole genome adequately, and (3) analytic methods that are statistically powerful and can be employed to identify unbiased genetic associations (Cantor et al., 2010).

One can employ a single-variant or multiple-variant GWAS when we assume the independence of tested variants from the rest. Single-SNP GWAS could be performed using the chi-square test, as multiple-SNP analysis could be performed using various regression models. Regression models give the possibility to include information about confounding factors, whilst other simple inference methods do not have this feature. Linear and logistic regression models in GWAS include information about covariates, effect sizes, genotype values for all individuals at SNP*s*, SNP effect sizes, the polygenic effect of other SNPs, the additive genetic variation of the phenotype, and standard genetic relationship.

Typically, in COVID-19 GWAS studies analysed, models are corrected for sex, age, and principal components (to remove the effect of population structure) (Dey et al., 2021; Wang et al., 2022). Additional information about comorbidities, laboratory test results, patients' clinical information, and genotyping array type are used as covariates in different study designs (Kuo et al., 2020; Zhu et al., 2021; Maes et al., 2022).

There is a consensus in COVID-19 HGI that GWAS studies should use a logistic regression association model including variant, age, age squared, sex, age multiplied by sex, 20 PCs, and study-specific variables as covariates. This requirement is applied to ensure the unified methodology for the proper interpretability of results shared through the COVID-19 HGI platform. As COVID-19 HGI embraces sharing scientific knowledge, summary statistics of GWAS analyses are available for other scientists to use in their research.

Tanimine et al. (2021) have applied a variable selection with forward-backward stepwise logistic regression to predict SNPs related to the risk of severe COVID-19 disease. However, as Ayers and Cordell (2010) noted, forward stepwise regression makes decisions in variable selection worse when the model becomes larger. Dite et al. (2021) have applied multivariable logistic regression while performing candidate variable selection by adding or removing additional candidate variables leaving only those with P<0.05. Moreover, Hu et al. (2021) proposed the super-variant concept (a set of alleles from multiple loci located anywhere in the genome) to find the association between them and the mortality of COVID-19 by applying logistic regression. The biggest issue in medical experiments is the insufficient data for computational methods. Furthermore, the number of individuals in different groups is usually unequal. To combat this problem, Grimaudo et al. (2021) applied multivariable logistic regression models to find the associations between the genotypes of mild and severe outcome patients and possible confounders (age, sex). The authors used only one genotype at a time to remove the bias of unbalanced data (different genotype distribution). The Firth logistic regression also combats the same problem and reduces slight sample bias in maximum likelihood estimation by penalising the likelihood and thus letting the separation occur during classification (Heinze and Schemper, 2002). Horowitz et al. (2020) applied Firth's logistic regression to determine associations between SNPs and seven defined phenotypes of COVID-19. The results of this analysis allowed to support the hypothesis that ACE2 levels influence COVID-19 risk.

Instead of using genotypes, Pairo-Castineira et al. (2021) tested genetic associations with patient status using a logistic regression model with a dose of gene alleles. Furthermore, Shelton et al. (2021) used logistic regression considering additive allele effects. In the simulation study performed by Setakis et al. (2008), authors have shown that methods based on simple logistic regression models perform well in all scenarios. In contrast, other methods (e.g. AdmixMap, Genomic Control) had issues with inflated false positive rates and had a low power when cases arose in only one subpopulation.

Ma et al. (2021) have applied this approach to conduct a gene-based association analysis to test the association between host genetic characteristics and the risk of developing COVID-19. The authors have identified multimarker aggregated effects and account for SNP p-values and linkage disequilibrium (LD) between SNPs by applying a multiple linear principal components regression approach. This technique achieves equivalent results as more sophisticated logistic regression models but with smaller computational resources.

Although regression models can detect associations between genotypes and disease status, they cannot detect causality or statistical coupling. Furthermore, a more advanced analysis could be conducted when we allow for interdependence between SNPs. Detailed information about the studies employing GWAS methods to model COVID-19 outcomes is summarised in Table 3.

Exome-wide association analysis (EWAS)

One of the GWAS variations, exome-wide association analysis (EWAS), differs from GWAS in sample and library preparations for sequencing. The aim is to capture only the exonic parts of the human genome. Kosmicki et al. (2020) applied EWAS to test whether there are significant associations between rare coding variants and COVID-19 outcomes (see Table 3). However, as the authors mentioned in their article, this method requires a large sample size. Thus the sample size of 7 million rare variants in around 20 thousand protein-coding genes was insufficient to reach a genome-wide significance. More studies with EWAS for COVID-19 modelling with larger sample sizes should be conducted to reach statistically significant results.

Phenome-wide association study (PheWAS)

A phenome-wide association study (PheWAS) aims to find the associations between SNPs (or other genetic features) and a variety of phenotypes (any trait, disease, or other) (Pendergrass et al., 2011). This approach is usually an additional step in GWAS, and it differs from GWAS in that the PheWAS association analysis starts with the specific DNA variant in order to find a possible phenotype. Although GWAS analysis is conducted reversely, searching for the SNPs associated with particular phenotypes. COVID-19 Host Genetics Initiative (2021) applied PheWAS to find out whether SNPs associated with other lung diseases could be related to the outcome of COVID-19. Verma et al. (2021) applied PheWAS separately for different ancestries and used phenotypes obtained from electronic health records (EHR). Authors applied either logistic or firth regression with adjustment for sex, age, age squared, and the first 20 principal components to test the association between SNPs and phenotypes. Associations were considered significant when the p-value was below 0.1 using the FDR correction (see Table 3).

4.2. Machine learning methods

Machine learning (ML) methods allow systems to automatically learn new experiences from the previous experience (Woolf, 2010). Although ML methods have become a common choice to solve various problems, scientists are just starting to apply them for disease prediction using patients' genomic data. Based on the review study (Kushwaha et al., 2020), a variety of ML methods have been applied for the COVID-19 pandemic, including regression, clustering, classification, transfer learning, ensemble methods, neural networks and deep learning, dimensionality reduction, reinforcement learning, word embeddings, and natural language processing. Mieth et al. (2016) have shown that machine learning methods (in this particular study - SVM) can identify genetic loci and increase the statistical power of GWAS. Next, we will review studies that have applied ML techniques to model COVID-19 using genomic patient data (summary presented in Table 4).

One of the most extensively applied ML methods is Random Forest (RF), based on stacked decision trees' votes (Ho Tin Kam, 1995). As Goldstein et al. (2011) noted, the RF model is suitable for studying genetic associations due to its ability to predict and present variable importance. Moreover, this algorithm, without any substantial difficulties, can handle thousands of observations and hundreds of thousands of predictors. Wang R.Y. et al. (2020) applied the RF model to the SNP number in the haplotype blocks to predict an individual's COVID-19 status and identify the essential haplotype blocks for COVID-19. The authors used an RF model with 248 trees in the forest and a random state parameter of 140. Even though the model was applied using insufficient data, it has reached an accuracy of 90%. These results suggest that incorporating the RF model on genomic data could get noteworthy results.

XGBoost (eXtreme Gradient Boosting) (Chen et al., 2016) is a gradient-boosting-based method that has outperformed other techniques for various sets of features in various settings. Similarly as RF, XGBoost is a tree ensemble method and is beneficial in genetic association studies because of their efficient handling of missing values, irrelevant and correlated variables, and they are computationally efficient to use. XGBoost was applied on the chromosomal-scale length variation (CSLV) (Toh and Brody, 2020) or laboratory test results (Wang F. et al., 2020) to predict which patients will develop a severe COVID-19 clinical outcome. However, this approach did not reach clinically applicable efficiency.

LASSO logistic regression is a standard ML algorithm used to solve binary classification tasks simultaneously, having a possibility to select the most significant features for model prediction. Moreover, it penalizes the regression coefficients and, in that way, allows more precisely finding the associated haplotypes, especially the rare ones (Biswas and Lin, 2012). Seven rare genetic variants of the Toll-like receptor gene associated with the outcome of COVID-19 in men were identified using LASSO logistic regression with a Boolean representation of genes on the X chromosome with rare variants (Fallerini et al., 2021). LASSO logistic regression with misclassification penalisation reduces the effect of unbalanced classes. Using this method, Fallerini et al. (2022) showed that additional genetic information (or, more precisely, Integrated Polygenic Score, IPGS) improves the ability to predict the severity of COVID-19.

Table 3. Methods applied in association analyses modelling COVID-19. All reviewed studies have used SNPs as an input for the models, except for exome-wide association analysis authors have used exomes. NA - information not available.

Modelling method		Publication	Sample size		Data source	Outcome
			N (cases/controls)	Comments		
GWAS	Logistic regression	COVID-19 Host Genetics Initiative, 2021	5.37 mln (70k/5.3 mln)	Critical illness due to COVID-19 (n=6,179 cases and n=1,483,780 controls), hospitalization due to COVID-19 (n=13,641 cases and 2,070,709 controls), and reported SARS-COV-2 infection (n=49,562 cases and n=1,770,206 controls)	Studies performed in 46 different laboratories	COVID-19 disease severity
		Severe Covid-19 GWAS Group, 2020	3,815 (1,610/2,205)	Cases – patients with severe COVID-19, controls – genotyped patients with other diseases (unknown COVID-19 status)	Seven hospitals in the Italian and Spanish epicentres	Disease severity (with mechanical ventilation or without)
		Pairo-Castineira et al., 2021	102,244+ (2,244/100,000+)	Cases from GenOMICC European cohort, controls – from UK Biobank, Generation Scotland (n=7,689) and 100,000 Genomes Project; n=1,675 individuals from the GenOMICC study and n=45,875 unrelated participants of European ancestry; An undefined number of individuals from 1000 Genomes Project	GenOMICC, ISARIC 4C, 1000 Genomes Project, UK Biobank	Having a disease/healthy individual
		Dey et al., 2021	12,389 (4,000/8,389)	Cases – severe COVID-19 patients (inpatients with a positive test), controls – non-severe COVID-19 patients	UK Biobank	Loci association with COVID-19 disease severity
		Hu et al., 2021	1,096 (292/804)	1,096 COVID-19 infected participants, of which 292 are deaths and 804 are survivors	UK Biobank	Association between SNPs and COVID-19-caused mortality/death
		Zhu et al., 2021	466 (466/0)	n=170 with mild COVID-19 symptoms, n=296 with severe COVID-19 symptoms	Wuhan Union Hospital	Association between SNPs and COVID-19 disease severity
		Verma et al., 2021	455,683 (NA/NA)	455,683 VAMVP participants, critical (n=35) and hospitalised (n=42) COVID-19 patients and controls. Two type GWAS: critical vs population, hospitalised vs population	EHR and genomic data from two biobanks: Veteran Affairs Million Veteran Program (VAMVP), United Kingdom Biobank (UKBB)	Association between SNPs and critical and hospitalised status of COVID-19
		Grimaudo et al., 2021	383 (383/0)	Mild or severe COVID-19 Sicilian patients	Laboratory for COVID-19 Surveillance for Western Sicily	SNP association with COVID-19 severity
		Dite et al., 2021	18,221 (1,713/16,508)	Cases – severe cases (inpatients), controls – non-severe cases (outpatients)	UK Biobank	COVID-19 disease severity (severe/non-severe)
		Horowitz et al., 2020	662,403 (11,356/651,047)	Cases – with COVID-19, controls – without COVID-19	AncestryDNA COVID-19 Research, Geisinger Health System, Penn MedicineBioBank, UK Biobank	Association between *ACE2* gene variants and risk of COVID-19 disease

Modelling method		Publication	Sample size		Data source	Outcome
		Kuo et al., 2020	322,948 (622/323,570)	Cases – with COVID-19, controls – without COVID-19	UK Biobank	Association between SNPs and COVID-19 severity
		Tanimine et al., 2021	230 (230/0)	Patients with COVID-19	3 Hospitals in Hiroshima, Japan	Associations between genotypes and risk of severe COVID-19 disease
		Shelton et al., 2021	114,240 (12,972/101,268)	Cases – with COVID-19, controls – without COVID-19	23andMe	Association between SNPs and COVID-19 risk
		Maes et al., 2022	528 (528/0)	n=308 critical, n=63 moderate, n=157 mild COVID-19 outcome	University Hospital of Londrina (HU) and the Emergency Rooms (ER) in Londrina, Paraná, Brazil	Associations between the sickness symptom complex (SSC) and COVID-19 and SNPs
		Wang D. et al., 2022	10,118 (NA/NA)	n=10,118 tested for COVID-19, n=1,265 with COVID-19, of those n=194 were fatal and n=1071 non-fatal	UK Biobank	Associations between Rab46 SNPs and COVID-19 fatality
	Multiple linear regression	Ma et al., 2021	680,128 (3,288/676,840)	1,610 cases and 2,205 controls from UK Biobank; 1,678 COVID-19 patients and 674,635 controls from COVID-19 Host Genetic Consortium	UK Biobank, COVID-19 Host Genetic Consortium	Association between SNPs and COVID-19 risk
Exome-wide association analysis		Kosmicki et al., 2020	543,213 (8,248/534,965)	n=8,248 had COVID-19, and among those n=2,085 (25.28%) were hospitalized and n=590 (7.15%) had severe disease	Geisinger Health System, Penn Medicine BioBank and UK Biobank	Disease susceptibility and disease severity
PheWAS		COVID-19 Host Genetics Initiative, 2021	5.37 mln (70k/5.3 mln)	Critical illness due to COVID-19 (n=6,179 cases and n=1,483,780 controls), hospitalization due to COVID-19 (n=13,641 cases and 2,070,709 controls), and reported SARS-COV-2 infection (n=49,562 cases and n=1,770,206 controls)	Studies performed in 46 different laboratories	Association between SNPs and defined COVID-19 phenotypes (severity)
		Verma et al., 2021	455,683 (NA/NA)	455,683 VAMVP participants, critical (n=35) and hospitalised (n=42) COVID-19 patients and controls. The numbers of cases and controls are presented for each studied phenotype separately	EHR and genomic data from two biobanks: Veteran Affairs Million Veteran Program (VAMVP), United Kingdom Biobank (UKBB)	Association between SNPs and defined COVID-19 phenotypes (severity)

Wu et al. (2009) have shown that penalised LASSO logistic regression can quickly and accurately identify predictors in situations with many uncertainties. Moreover, it identifies associations previously missed by other methods.

Ensemble learning in machine learning is often used to obtain better prediction performance using an ensemble of a finite number of selected ML models. Carapito et al., 2021 employed an ML ensemble of 7 models (LASSO regression, Ridge regression, Support Vector Machines (SVM), quantum SVM, XGBoost, RF, Deep Artificial Neural Network) to predict the severity of COVID-19 using data from patient RNA sequencing. This ensemble, with high accuracy (91%) and efficiency (AUC = 0.94), selects five genes (*ADAM9*, *RAB10*, *MCEMP1*, *MS4A4A*, and *GCLM*) out of 600 as the most important ones deciding the disease severity. The positive unlabeled machine learning model and a stable feature learning framework RubricOE (learning rubric for multi-omics and genetic epidemiology) is an ML ensemble that helped to determine the genomic factors driving the severity of COVID-19 (Dey et al., 2021). The authors disclose that a combination of human genomic and clinical data improves the accuracy of severe COVID-19 cases prediction.

Artificial neural networks

Santus et al. (2021) provided four main research and development areas where artificial intelligence might be applied to combat the COVID-19 pandemic: (1) triage, diagnosis, and risk of mortality/severity prediction; (2) drug repurposing and development (modelling virus-host interactions); (3) pharmacogenomics and vaccines (genetic markers identification, infection susceptibility prediction); and (4) mining of the medical literature (examining the quality of results in countless scientific studies).

The ability of artificial neural networks (ANNs) to process huge amounts of data, learn complex features and nonlinear and interaction effects, makes them suitable for genomic data analysis. They can model complex relations between traits and genetic host features without the need to specify all the possible interactions between those features. A multilayer perceptron classifier with two hidden layers and a Rectified Linear Unit (ReLU) activation function between them was applied to the number of SNPs in a specific haplotype block to predict an individual’s COVID-19 status (Wang R.Y. et al., 2020). The model reached high accuracy and precision, suggesting ANNs to be an appropriate choice for analysing the genomic data.

4.3. Other statistical methods

Scientists often select genes known to be associated with some specific diseases or phenotypes and then try to predict the effect of changes in these genes on the translated protein. As summarised in **Table 3** and **4**, many genes have been described as associated with COVID-19 by employing association analysis and machine learning methods. Next, we review the studies that model COVID-19 using a single gene and apply the Sherlock-based integrative genomics analysis (summary in **Table 5**).

Single-gene analysis

Due to the lack of computational and financial resources, selecting a limited number of genes and analysing only their genetic variants is a common practice. The main interest here is whether those genetic variants influence the structure and function of translated proteins.

Table 4. Machine learning techniques applied to model COVID-19 outcome. Abbreviation NA here means information not available.

Modelling method	Publication	Sample size		Laboratory methods	Data source	Data type	Outcome
		N (cases/controls)	Comments				
Artificial neural network Random forest	Wang R.Y. et al., 2020	673 (NA/NA)	NA	SNP genotyping assay	From The Personal Genome Project dataset (Ball et al., 2012)	SNPs	Disease severity ((I) COVID-19 infected, (II) hospitalised, and (III) severe conditions)
XGBoost	Toh and Brody, 2020	~2,000 (981/~981)	Cases – patients with COVID-19, controls – similar age individuals from the general UK Biobank population	DNA microarray	UK Biobank	Chromosomal-scale length variation data	Severe/not severe COVID-19
	Wang F. et al., 2020	332 (332/0)	Patients with different COVID-19 severity: asymptomatic, mild, moderate, severe and critically ill	Whole-genome sequencing	The same study	SNPs	COVID-19 disease severity (from 1 to 5)
LASSO logistic regression	Fallerini et al., 2021	156 (79/77)	Cases – male patients with COVID-19 and air ventilation, controls – asymptomatic male patients with COVID-19	SNP genotyping assay	Italian GEN-COVID (Daga et al., 2021)	SNPs	Feature importance
	Fallerini et al., 2022	4,591 (2,944/1,647)	Cases – patients with severe COVID-19 outcome, controls – patients with mild COVID-19 outcome	Whole-exome sequencing	6 data sources*	SNPs	Predicting the COVID-19 phenotype from Boolean features of protein-changing genetic variants with correction of age and sex
Ensemble of multiple ML models	Carapito et al., 2021	72 (47/25)	n=47 critical (C) COVID-19 patients, n=25 non-critical (NC) COVID-19 patients	Whole-transcriptome RNA-seq	University hospital network in northeast France (Alsace)	Gene expression	Classification of NC versus C patients, finding a gene signature
Positive-unlabeled learning algorithms coupled with RubricOE	Dey et al., 2021	12,389 (4,000/8,389)	Cases – severe COVID-19 patients (inpatients with a positive test), controls – non-severe COVID-19 patients	SNP genotyping assay	UK Biobank	SNPs	SNP association with COVID-19 disease severity

* GEN-COVID (Italy); The genetic predisposition to severe COVID-19 (Sweden); German COVID-19 OMICS Initiative (Germany); Quebec COVID-19 Biobank and Swedish Biobank (Canada-Sweden); Biobanque Québécoise de la Covid-19 (Canada); GenOMICC/ISARIC4C (UK)

The SNPs in the coding regions might change the function or structure of the coded proteins. Protein prediction algorithms, such as Polyphen-2 HumDiv, Poplyphen HumVar, Sorting Intolerant from Tolerant (SIFT), logistic regression test scores, MutationTaster, DUET program (Pires et al., 2014), and GROMACS software (Abraham et al., 2015), have been applied to determine the possible structural changes of ACE2 human receptor caused by specific genetic variants (Guo et al., 2020; Benetti et al., 2020).

One of the main steps in entering the host organism for the SARS-CoV-2 virus is binding to the cell surface through the interaction of the viral S protein and the human ACE2 receptor. Therefore, the ACE2 human cell receptor is widely studied. Suryamohan et al. (2020) analysed *ACE2* genotypes and calculated the fixation index (Fst) to determine the genetic variation in the *ACE2* gene, resulting in a different number of significant genetic variants for different datasets.

The Genome Aggregation Database (GnomAD) (Cummings et al., 2020) with aggregated exome and genome sequencing data is a common choice for selecting genomic data from COVID-19 patients. *ACE2* (Gibson et al., 2020; MacGowan and Barton, 2020) and *TMPRSS2* (Hou Yuan et al., 2020) coding variants were obtained from this database for the theoretical modelling of rare *ACE2* coding variants and the effect of *ACE2* gene variants on binding to viral S-protein.

The linear regression model was applied to discover nine relatively common variants and six missense variants of *TMPRSS2* that can negatively affect the activity of this protease, while only one variant significantly decreases the risk of COVID-19 (Hou Yuan et al., 2020; Monticelli et al., 2021). Wulandari et al. (2021) have selected a particular *TMPRSS2* polymorphism, p.Val160Met, possibly associated with the severity of COVID-19 disease. However, the linear-by-linear chi-square association test did not find a significant association, possibly due to the small sample size (n=95).

One variant of the *IFITM3 gene and its association with COVID-19 severity was studied using simple statistical tests (Zhang et al., 2020). While three out of* nine selected *LZTFL1* SNPs were shown to be associated with COVID-19 severity in the Latvian population by the logistic regression model (with age and sex as covariates and correction for population stratification) (Rescenko et al., 2021). Secolin et al. (2021) have found three SNPs of the *HLA* gene that increase the risk of severe COVID-19 in a Brazilian population.

Sherlock-based integrative genomics analysis

Sherlock-based integrative genomics analysis is based on the Bayesian inference algorithm (He et al., 2013). It searches for SNPs associated with gene expression (called eSNPs), then the possible association of the phenotype is estimated. The Bayes factor (LBF) logarithm is calculated for each pair of SNPs and summed to get the LBF for each gene. The positive LBF will be given for those eSNPs that showed a statistically significant association with the phenotype. Negative LBF will be assigned for eSNPs without significant association with the studied phenotype. Then the p-value of the LBF for each gene is computed using simulation analysis. This method estimated seven genes and their variants associated with COVID-19 in 49 types of human tissues (Ma et al., 2021).

Table 5. Other statistical methods applied to model COVID-19. All studies used SNPs as an input to the models. Abbreviation NA here means information not available.

Modelling method	Publication	Sample size		Data source	Outcome
		N (cases/controls)	Comments		
Sherlock	Ma et al., 2021	680,128 (3,288/676,840)	1,610 cases and 2,205 controls from UK Biobank; 1,678 COVID-19 patients and 674,635 controls from COVID-19 Host Genetic Consortium	UK Biobank, COVID-19 Host Genetic Consortium	Association between SNPs and COVID-19
Other statistical methods	Suryamohan et al., 2020	60,164 (NA/NA)	n=2,381 from the 1000 Genomes Project Phase 3 and n=57,783 female individuals from gnomAD	1000 Genomes Project, gnomAD	Estimate of genetic variation
	Benetti et al., 2020	389 (131/258)	131 individuals belonging to the GEN-COVID MULTICENTER STUDY (Giliberti et al., 2020), controls consist of 258 Italian individuals	gnomAD	Variant effect on protein function prediction
	Guo et al., 2020	141,456 (NA/NA)	n=125,748 from whole-exome sequencing and n=15,708 – whole-genome sequencing	gnomAD	Variant effect on protein function prediction
	Gibson et al., 2020	141,456 (NA/NA)	n=76,702 of male and n= 64,754 of female participants' exomes/genomes	gnomAD	Variant effect on protein function prediction
	MacGowan and Barton, 2020	NA	NA	gnomAD	Variant effect on protein function prediction
	Hou Yuan et al., 2020	~81,000 (NA/NA)	437 non-synonymous single-nucleotide variants from 81 thousand genomes	GnomAD, Exome Sequencing Project (ESP), 1000 Genomes Project	Association between *ACE2* and *TMPRSS2* DNA polymorphisms with COVID-19 severity
	Wulandari et al., 2021	95 (NA/NA)	62 patients with moderate and severe COVID-19 from Dr Soetomo General Academic Hospital, 33 patients with asymptomatic or mild symptoms from Indrapura KOGABWILHAN II Hospital	Dr Soetomo General Hospital and Indrapura Field Hospital (Surabaya, Indonesia)	Correlation between a genetic variant within the human *TMPRSS2* gene and COVID-19 severity and viral load
	Monticelli et al., 2021	1,177 (NA/NA)	Patients affected with COVID-19 in Italy	GEN-COVID Multicenter Study	Correlation between protein variants with the clinical features of COVID-19 patients
	Secolin et al., 2021	386 (NA/NA)	NA	BIPMed (www.bipmed.org) and ABraOM(abraom.ib.usp.br) datasets	Genetic variation in COVID-19-related genes in the Brazilian population
	Zhang et al., 2020	80 (NA/NA)	Patients with mild (56) and severe (24) COVID-19. Not a case-control study	Beijing Youan Hospital, Capital Medical University, Beijing	Association between rs12252 SNP and COVID-19 severity
	Rescenko et al., 2021	2,692 (475/2,217)	Individuals from the Latvian population	Genome Database of Latvian Population	Association between SNPs and increased risk of COVID-19 and hospitalisation status

5. Discussion

According to current scientific knowledge, seven zoonotic coronaviruses are capable of causing infections in humans. Although the disorder usually manifests with mild features, recently, the world experienced two global outbreaks associated with many lethal outcomes (SARS-CoV in 2002 and MERS-CoV in 2012). SARS-CoV-2 has a much wider spread and has affected more people than previous entities. Demographic characteristics (age, sex, ethnicity), comorbidities, some clinical symptoms, specific laboratory test findings, and diet/lifestyle predetermine a patient's risk of developing the more severe course of the disease. Morbidity and mortality due to COVID-19 rise dramatically with age and co-existing health conditions. Still, even very young and otherwise healthy patients can unpredictably succumb to this disease.

Previous studies of SARS and MERS infections have provided insight into possible functional pathways of disease development, yet the prior experience cannot be directly applied to this new virus entity. The mechanisms underlying the spectrum of COVID-19 characteristics remain largely unknown. Obviously, viral and human (host) genetic factors are suspected to play a significant role in pathogenesis, and there is already considerable evidence that the genetic characteristics of human genomes have a clear association with the severity of COVID-19 (Zhang et al., 2020; Wang et al., 2020; COVID-19 Host Genetics Initiative 2021). The data from several studies suggest that host genetic factors determine the 'predicted COVID-19' phenotype in 50% of cases (Williams et al., 2020).

Since the beginning of an outbreak, many human genome variants (see Supplementary **Tables S1** and **S2**) have been linked with the increased or decreased risk of developing rough COVID-19. About 70 % of identified variants, namely in *ACE2, TMPRSS2* and *TLL-1* genes associated with the virus attachment and entry, decrease the risk of COVID-19, while almost all identified variants associated with the immune response in 6p21.3 (*HLA* genes*)*, 9q34 (*ABO* genes), 3p21.31 (*SLC6A20, LZTFL1, FYCO1, CXCR6, XCR1, CCR9* genes*)*, and 12q24.13 (*OAS1/2/3* genes) genetic loci as well as in *IFIH1* and *IL6* genes increase the risk of severe COVID-19.

When reliable genetic associations for various phenotypes are known, scientists are faced with the next big challenge: interpreting these associations in a biological and genomic context. Prediction of disease characteristics using computational methods is one of the areas where big healthcare data and computational methods can merge to provide potentially more accurate diagnoses for patients. Naturally, it is essential to understand the biology behind SARS-CoV-2 infection to interpret the results of applied computations correctly. Therefore, the system biology approach offers the most appropriate way to reach the purpose and the tight collaboration of biologists, physicians, mathematicians and programmers is highly appreciable and recommended.

A wide variety of experimental and computational methods have been applied to find associations of human genetic characteristics with the susceptibility and outcomes of COVID-19. Whole-genome/exome genotyping data or specific genomic variants are used to estimate the association. These methods include conventional methods such as genome-wide association study (GWAS), statistical methods like linear or logistic regression, and more sophisticated machine learning methods including Random Forest, XGBoost, LASSO logistic regression, artificial neural networks, and others.

One of the most significant drawbacks of conventional methods used for modelling the disease phenotypes is that they do not analyse the causality relationships between genetic markers and the phenotype of interest (Sun et al., 2020). Usually, these methods use one type of data (e.g. genotyping data) to look for differences between individuals in the population. As association studies mainly apply a series of single-SNP statistical tests, they miss the correlations between SNPs and possibly discard relevant SNPs with a low effect on the phenotype (Yang et al., 2010). While discussing the association studies of genetic variants with COVID-19 susceptibility and possible outcomes, we must remember that additional factors such as selected groups of subjects, non-genetic factors, and other co-variables such as age and gender can impact the results. In that way, complex diseases resulting from the effect of multiple genes cannot be efficiently studied. Models using additional information, such as clinical data or more complex approaches that evaluate interactions between several places in the genome, could help see the whole perspective. Therefore, various novel methods, such as machine learning for genomic data analysis and association studies, are being applied. The application of advanced computational methods (e.g., machine learning methods, artificial neural networks) for predicting COVID-19 could potentially suggest new insights into the previously unseen. As Stoeger and Amaral (2020) suggest, scientists are prone to study the formerly known genes associated with the studied object, leaving behind the potential new host genomic candidates. Furthermore, conventional methods are insufficient to study complex diseases associated with more than one gene variant/locus. The use of novel methods could help to find the associations mentioned earlier.

Sun et al. (2021) reviewed statistical modelling and machine learning methods as a replacement for the GWAS and concluded that one method could not solve all problems; therefore, conventional GWAS methods will still be used in the future. However, the additional information provided using novel methods adds value to scientific knowledge and allows us to see the bigger picture. A combination of GWAS and machine learning algorithms might be a promising solution for this issue.

After a comprehensive review of artificial intelligence (AI) implications in COVID-19 studies, Rasheed et al. (2021) observed that AI methods are applied to estimate the disease severity in patients from chest CT or X-ray images clinical or time-series data. In addition, Monte Carlo-based simulations, hidden Markov models, and neural networks are used to predict vaccine targets. Díez Díaz et al. (2021) have developed a machine learning methodology to study multi-SNP associations by incorporating the GWAS data and sophisticated ML algorithms - genetic algorithms together with support vector machines. However, there is an insufficient number of studies applying advanced machine learning/AI methods to human genomic data. Further investigation is required to precisely understand the mechanisms behind the disease genetics in the human organism.

Mendelian randomisation studies are being enrolled to determine causal relationships between risk factors and diseases. Namely, Wu et al. (2021) have applied an integrative multi-omics approach by combining the cross-methylome omnibus (CMO) method with association analysis with S-PrediXcan and fine-mapping of gene sets strategy to discover putative causal genes for COVID-19. However, causal relationships between factors and COVID-19 are understudied, and research is only beginning to emerge, suggesting an exciting topic for future studies.

As far as sufficient precision for medical decisions is concerned, the critical factor of not reaching adequate accuracy is too little data. Fortunately, the amount of data is constantly expanding and waiting to be harnessed for the well-being of humanity. At the current stage, many studies are seeking to recruit a variety of available data (computer tomography (CT)/X-ray scans, clinical data, epidemiological data, various omics data, and others) and different methods to combat COVID-19. Additional data can lead to a more accurate disease prognosis for patients by employing computational approaches to test research hypotheses or add additional information on the topic. Nevertheless, after analysing 169 studies with COVID-19 prediction models, Wynants et al. (2020) concluded that even though scientists are making models to predict the outcome/severity of COVID-19 or the overall risk of developing this disease, the quality of those methods is poor, and the operation of them in practice might be unreliable. Large-scale studies with larger cohorts of individuals are required to elucidate the host response to SARS-CoV-2, and more COVID-19 prediction studies should be enrolled correctly validating the created methods using enough real-life data and/or data from other studies. An increasing number of the collected data in the healthcare system opens up opportunities to use them for scientific and medical purposes.

An urgency in obtaining insights into COVID-19 pathogenesis and highly variable clinical manifestations of SARS-CoV-2 infection for improved management, better patient outcomes and disease prevention is critical for the scientific and healthcare communities. By using this knowledge, we could protect the most vulnerable people and prevent the worst outcomes of COVID-19.

6. Conclusions

The COVID-19 pandemic is a global crisis that creates severe disruptions in the economy and health system. Insights into better understanding and treatment of COVID-19 are desperately needed. Given the importance and urgency of obtaining these insights, the scientific community must come together around this shared purpose.

The collection of massive genomic and health data followed by comprehensive biostatistics/bioinformatics analysis will enable the identification of genomic factors that influence the characteristics of the disease. Learning the genetic determinants of susceptibility, severity, and outcomes of COVID-19 could contribute to the translation of the findings into patient care and disease prevention, help generate hypotheses for drug repurposing, identify individuals at unusually high or low risk, and contribute to global knowledge of the biology of SARS-CoV-2 infection and disease.

References

Abraham, M .J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., Lindahl, E. (2015). GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, *1*, pp.19-25. https://doi.org/10.1016/j.softx.2015.06.001

Argenziano, M. G., Bruce, S. L., Slater, C. L., Tiao, J. R., Baldwin, M. R., Barr, R. G., Chang, B. P., Chau, K. H., Choi, J. J., Gavin, N., Goyal, P. (2020). Characterisation and clinical course of 1000 patients with coronavirus disease 2019 in New York: retrospective case series. *Bmj*, **369**. https://doi.org/10.1136/bmj.m1996

Rosanna, A., Paraboschi, E. M., Mantovani, A., Duga, S. (2020). ACE2 and TMPRSS2 Variants and Expression as Candidates to Sex and Country Differences in COVID-19 Severity in Italy. Preprint. *Genetic and Genomic Medicine*. https://doi.org/10.1101/2020.03.30.20047878

Ayers, K. L., Cordell, H. J. (2010). SNP selection in genome- wide and candidate gene studies via penalized logistic regression. *Genetic epidemiology*, *34*(8), pp.879-891. https://doi.org/10.1002/gepi.20543

Ball, M. P., Thakuria, J. V., Zaranek, A. W., Clegg, T., Rosenbaum, A. M., Wu, X., Angrist, M., Bhak, J., Bobe, J., Callow, M. J., Cano, C. (2012). A public resource facilitating clinical use of genomes. *Proceedings of the National Academy of Sciences* **109**(30), pp.11920-11927. https://doi.org/10.1073/pnas.1201904109

Benetti, E., Tita, R., Spiga, O. et al. (2020). ACE2 gene variants may underlie interindividual variability and susceptibility to COVID-19 in the Italian population. *Eur J Hum Genet* **28**, 1602–1614. https://doi.org/10.1038/s41431-020-0691-z

Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), pp.289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Biswas, S., Lin, S. (2012). Logistic Bayesian LASSO for identifying association with rare haplotypes and application to age‐ related macular degeneration. Biometrics, 68(2), pp.587-597. https://doi.org/10.1111/j.1541-0420.2011.01680.x

Blanco-Melo, D., Nilsson-Payant, B., Liu, W. C., Møller, R., Panis, M., Sachs, D., Albrecht, R. (2020). SARS-CoV-2 launches a unique transcriptional signature from in vitro, ex vivo, and in vivo systems. Preprint. *bioRxiv*. https://doi.org/10.1016/j.cell.2020.04.026

Cantor, R. M., Lange, K., Sinsheimer, J. S. (2010). Prioritizing GWAS results: A review of statistical methods and recommendations for their application. American journal of human genetics, 86(1), 6–22. https://doi.org/10.1016/j.ajhg.2009.11.017

Carapito, R., Li, R., Helms, J., Carapito, C., Gujja, S., Rolli, V., Guimaraes, R., Malagon-Lopez, J., Spinnhirny, P., Lederle, A., Mohseninia, R. (2021). Identification of driver genes for critical forms of COVID-19 in a deeply phenotyped young patient cohort. *Science Translational Medicine*, p.eabj7521. https://doi.org/10.1126/scitranslmed.abj7521

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**(1), pp.s13742-015. https://doi.org/10.1186/s13742-015-0047-8

Chen, T., Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (New York, NY, USA, 2016), KDD '16, ACM* (pp. 785-794). https://doi.org/10.1145/2939672.2939785

COVID-19 Host Genetics Initiative (2021). Mapping the human genetic architecture of COVID-19. *Nature*. https://doi.org/10.1038/s41586-021-03767-x

Coony, E. (2020). Long after the fire of a COVID-19 infection, mental and neurological effects can still smolder. *STAT. URL https://www. statnews. com/2020/08/12/after-covid19-mental-neurological-effects-smolder/* (Accessed on June 9, 2021).

Cummings, B. B., Karczewski, K. J., Kosmicki, J. A., Seaby, E. G., … Watts, N. A. (2020). Transcript expression-aware annotation improves rare variant interpretation. *Nature*, 581(7809), 452–458. https://doi.org/10.1038/s41586-020-2329-2

Daga, S., Fallerini, C., Baldassarri, M., Fava, F., Valentino, F., Doddato, G., Benetti, E., Furini, S., Giliberti, A., Tita, R., Amitrano, S. (2021). Employing a systematic approach to biobanking and analysing clinical and genetic data for advancing COVID-19 research. *European Journal of Human Genetics*, pp.1-15. https://doi.org/10.1038/s41431-020-00793-7

de Leeuw, C.A., Mooij, J.M., Heskes, T., Posthuma, D. (2015). MAGMA: generalised gene-set analysis of GWAS data. *PLoS computational biology* **11**(4), p.e1004219

Dey, S., Bose, A., Chakraborty, P., Ghalwash, M., Saenz, A.G., Ultro, F., Kenney, N.G., Hu, J., Parida, L., Sow, D. (2021). Impact of Clinical and Genomic Factors on SARS-CoV2 Disease Severity. Preprint. *medRxiv*. https://doi.org/10.1101/2021.03.15.21253549

Díez Díaz, F., Sánchez Lasheras, F., Moreno, V., Moratalla-Navarro, F., Molina De La Torre, A. J., Martín Sánchez, V. (2021). GASVeM: A New Machine Learning Methodology for Multi-SNP Analysis of GWAS Data Based on Genetic Algorithms and Support Vector Machines. *Mathematics* **9**(6), p.654. https://doi.org/10.3390/math9060654

Ding, S., Liang, T. J. (2020). Is SARS-CoV-2 also an enteric pathogen with potential fecal–oral transmission? A COVID-19 virological and clinical review. *Gastroenterology*, **159**(1), pp.53-61. https://doi.org/10.1053/j.gastro.2020.04.052

Dite, G.S., Murphy, N.M., Allman, R. (2021). An integrated clinical and genetic model for predicting risk of severe COVID-19: A population-based case–control study. *PloS one* **16**(2), p.e0247205. https://doi.org/10.1371/journal.pone.0247205

Ellinghaus, D., Degenhardt, F., Bujanda, L., Buti, M., Albillos, A., Invernizzi, P., Fernández, J. et al. (2020). The ABO Blood Group Locus and a Chromosome 3 Gene Cluster Associate with SARS-CoV-2 Respiratory Failure in an Italian-Spanish Genome-Wide Association Analysis. Preprint. *Infectious Diseases (except HIV/AIDS)*. https://doi.org/10.1101/2020.05.31.20114991

Fallerini, C., Daga, S., Mantovani, S., Benetti, E., Picchiotti, N., Francisci, D., Paciosi, F., Schiaroli, E., Baldassarri, M., Fava, F., Palmieri, M. (2021). Association of Toll-like receptor 7 variants with life-threatening COVID-19 disease in males: findings from a nested case-control study. *Elife* **10**, p.e67569. https://doi.org/10.7554/eLife.67569

Fallerini, C., Picchiotti, N., Baldassarri, M., Zguro, K., Daga, S., Fava, F., Benetti, E., Amitrano, S., Bruttini, M., Palmieri, M., Croci, S. (2022). Common, low-frequency, rare, and ultra-rare coding variants contribute to COVID-19 severity. *Human genetics*, **141**(1), pp.147-173. https://doi.org/10.1007/s00439-021-02397-7

Fehr, A. R., Perlman, S. (2015). Coronaviruses: An Overview of Their Replication and Pathogenesis. In Coronaviruses, edited by Helena Jane Maier, Erica Bickerton, and Paul Britton, 1282:1–23. *Methods in Molecular Biology*. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4939-2438-7

Fung, T. S., Liu, D. X. (2019). Human coronavirus: host-pathogen interaction. Annual review of microbiology, 73, pp.529-557. https://doi.org/10.1146/annurev-micro-020518-115759

Gao, X., Starmer, J., Martin, E. R. (2008). A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* ***32***(4), pp.361-369. https://doi.org/10.1002/gepi.20310

Gao, Y. D., Ding, M., Dong, X., Zhang, J. J., Kursat Azkur, A., Azkur, D., Gan, H., Sun, Y. L., Fu, W., Li, W., Liang, H. L. (2021). Risk factors for severe and critically ill COVID- 19 patients: a review. *Allergy* **76**(2), pp.428-455. https://doi.org/10.1111/all.14657

Gaziano, J. M., Concato, J., Brophy, M., Fiore, L., Pyarajan, S., Breeling, J., Whitbourne, S., Deen, J., Shannon, C., Humphries, D., Guarino, P. (2016). Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *Journal of clinical epidemiology*, *70*, pp.214-223. https://doi.org/10.1016/j.jclinepi.2015.09.016

Gibson, W. T., Evans, D. M., An, J., Jones, S. J. M. (2020). ACE 2 Coding Variants: A Potential X-Linked Risk Factor for COVID-19 Disease. Preprint. *Bioinformatics*. https://doi.org/10.1101/2020.04.05.026633

Giliberti, A., Benetti, E., Emiliozzi, A., Valentino, F., Bergantini, L., Fallerini, C., Anedda, F., Amitrano, S., Conticini, E., Tita, R., d'Alessandro, M. (2020). Clinical and molecular characterisation of COVID-19 hospitalised patients. *PLoS One*, **15**(11), p.e0242534. https://doi.org/10.1101/2020.05.22.20108845

Goldstein, B. A., Polley, E. C., Briggs, F. B. (2011). Random forests for genetic association studies. Statistical applications in genetics and molecular biology, 10(1). https://doi.org/10.2202%2F1544-6115.1691

Grimaudo, S., Amodio, E., Pipitone, R.M., Maida, C.M., Pizzo, S., Prestileo, T., Tramuto, F., Sardina, D., Vitale, F., Casuccio, A., Craxì, A. (2021). PNPLA3 and TLL-1 Polymorphisms as Potential Predictors of Disease Severity in Patients With COVID-19. *Frontiers in Cell and Developmental Biology* **9**, p.1589. https://doi.org/10.3389/fcell.2021.627914

Grimm, D.G., Roqueiro, D., Salomé, P. A., Kleeberger, S., Greshake, B., Zhu, W., Liu, C., Lippert, C., Stegle, O., Schölkopf, B., Weigel, D. (2017). easyGWAS: a cloud-based platform for comparing the results of genome-wide association studies. *The Plant Cell*, **29**(1), pp.5-19. https://doi.org/10.1105/tpc.16.00551

Guo, X., Chen, Z., Xia, Y., Lin, W., Li, H. (2020). Investigation of the Genetic Variation in ACE2 on the Structural Recognition by the Novel Coronavirus (SARS-CoV-2). *Journal of Translational Medicine* **18** (1): 321. https://doi.org/10.1186/s12967-020-02486-7

Gustine, J. N., Jones, D. (2021). Immunopathology of Hyperinflammation in COVID-19. *The American Journal of Pathology* **191** (1): 4–17. https://doi.org/10.1016/j.ajpath.2020.08.009

He, X., Fuller, C.K., Song, Y., Meng, Q., Zhang, B., Yang, X., Li, H. (2013). Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS. *The American Journal of Human Genetics* **92**(5), pp.667-680. https://doi.org/10.1016/j.ajhg.2013.03.022

Heinze, G., Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in medicine*, *21*(16), pp.2409-2419. https://doi.org/10.1002/sim.1047

Ho, T. K. (1995). Random Decision Forests. *Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.*

Horowitz, J. E., Kosmicki, J. A., Damask, A., Sharma, D., Roberts, G. H. L., Justice, A. E., Banerjee, N. et al. (2020). Common Genetic Variants Identify Targets for COVID-19 and

Individuals at High Risk of Severe Disease. Preprint. *Genetic and Genomic Medicine*. https://doi.org/10.1101/2020.12.14.20248176

Hou, Y., Zhao, J., Martin, W., Kallianpur, A., Chung, M. K., Jehi, L., Sharifi, N., Erzurum, S., Eng, C., Cheng, F. (2020). New Insights into Genetic Susceptibility of COVID-19: An ACE2 and TMPRSS2 Polymorphism Analysis. *BMC Medicine* **18** (1): 216. https://doi.org/10.1186/s12916-020-01673-z

Hu, J., Li, C., Wang, S., Li, T., Zhang, H. (2021). Genetic variants are identified to increase risk of COVID-19 related mortality from UK Biobank data. *Human genomics* **15**(1), pp.1-10. https://doi.org/10.1186/s40246-021-00306-7

Huang, Y, Yang, C., Xu, X., Xu, W., Liu, S. (2020). Structural and Functional Properties of SARS-CoV-2 Spike Protein: Potential Antivirus Drug Development for COVID-19. *Acta Pharmacologica Sinica* **41** (9): 1141–49. https://doi.org/10.1038/s41401-020-0485-4

Huffman, J., Butler-Laporte, G., Khan, A., Drivas, T.G., Peloso, G. M., Nakanishi, T., Verma, A., Kiryluk, K., Richards, J. B., Zeberg, H. (2021). Alternative splicing of OAS1 alters the risk for severe COVID-19. medRxiv. https://doi.org/10.1101/2021.03.20.21254005

Kaler, A. S., Purcell, L. C. (2019). Estimation of a significance threshold for genome-wide association studies. *BMC genomics* **20**(1), pp.1-8. https://doi.org/10.1186/s12864-019-5992-7

Kang, S., Yang, M., Hong, Z., Zhang, L., Huang, Z., Chen, He, S. et al. (2020). Crystal Structure of SARS-CoV-2 Nucleocapsid Protein RNA Binding Domain Reveals Potential Unique Drug Targeting Sites. *Acta Pharmaceutica Sinica B* **10** (7): 1228–38. https://doi.org/10.1016/j.apsb.2020.04.009

Kenney, A. D., Dowdle, J. A., Bozzacco, L., McMichael, T. M., St. Gelais, C., Panfil, A. R., Sun, Y., Schlesinger, L. S., Anderson, M. Z., Green, P. L., López, C. B. (2017). Human genetic determinants of viral diseases. *Annual review of genetics* **51**, pp.241-263. https://doi.org/10.1146/annurev-genet-120116-023425

Kosmicki, J. A., Horowitz, J. E., Banerjee, N. et al. (2020). A catalog of associations between rare coding variants and COVID-19 outcomes. *medRxiv*. https://doi.org/10.1101/2020.10.28.20221804

Kundu, R., Das, R., Geem, Z. W., Han, G. T., Sarkar, R. (2021). Pneumonia detection in chest X-ray images using an ensemble of deep learning models. *PloS one*, **16**(9), p.e0256630. https://doi.org/10.1371/journal.pone.0256630

Kuo, C. L., Pilling, L. C., Atkins, J. L., Masoli, J. A., Delgado, J., Kuchel, G. A., Melzer, D. (2020). APOE e4 genotype predicts severe COVID-19 in the UK Biobank community cohort. *The Journals of Gerontology: Series A* **75**(11), pp.2231-2232. https://doi.org/10.1093/gerona/glaa131

Kurki, S. N., Kantonen, J., Kaivola, K., Hokkanen, L., Mäyränpää, M. I., Puttonen, H., Martola, J., Pöyhönen, M., Kero, M., Tuimala, J., Carpén, O. (2021). APOE ε4 associates with increased risk of severe COVID-19, cerebral microhaemorrhages and post-COVID mental fatigue: a Finnish biobank, autopsy and clinical study. *Acta neuropathologica communications* **9**(1), pp.1-13. https://doi.org/10.1186/s40478-021-01302-7

Kushwaha, S., Bahl, S., Bagha, A. K., Parmar, K. S., Javaid, M., Haleem, A., Singh, R .P. (2020). Significant applications of machine learning for COVID-19 pandemic. *Journal of Industrial Integration and Management* **5**(4). https://doi.org/10.1142/S2424862220500268

Ma, Y., Huang, Y., Zhao, S., Yao, Y., Zhang, Y., Qu, J., Wu, N., Su, Y. (2021). Integrative Genomics Analysis Reveals a 21q22.11 Locus Contributing Risk to COVID-19. *Human Molecular Genetics*, May, ddab125. https://doi.org/10.1093/hmg/ddab125

MacGowan, S. A., Barton, G. J. (2020). Missense Variants in ACE2 Are Predicted to Encourage and Inhibit Interaction with SARS-CoV-2 Spike and Contribute to Genetic Risk in COVID-19. Preprint. *Genetics*. https://doi.org/10.1101/2020.05.03.074781

Maes, M., Tedesco Junior, W. L. D., Lozovoy, M. A. B., Mori, M. T. E., Danelli, T., Almeida, E. R. D. D., Tejo, A. M., Tano, Z. N., Reiche, E. M. V., Simão, A. N. C. (2022). In COVID-19, NLRP3 inflammasome genetic variants are associated with critical disease, and these effects are partly mediated by the sickness symptom complex: a nomothetic network approach. *Molecular Psychiatry*, pp.1-11. https://doi.org/10.1038/s41380-021-01431-4

Maiti, A. K. (2020). The African-American population with a low allele frequency of SNP rs1990760 (T allele) in IFIH1 predicts less IFN-beta expression and potential vulnerability to COVID-19 infection. *Immunogenetics*, **72**(6), pp.387-391. https://doi.org/10.1007/s00251-020-01174-6

Mieth, B., Kloft, M., Rodríguez, J. A., Sonnenburg, S., Vobruba, R., Morcillo-Suárez, C., Farré, X., Marigorta, U. M., Fehr, E., Dickhaus, T., Blanchard, G. (2016). Combining multiple hypothesis testing with machine learning increases the statistical power of genome-wide association studies. Scientific reports, 6(1), pp.1-14. https://doi.org/10.1038/srep36671

Monticelli, M., Hay Mele, B., Benetti, E., Fallerini, C., Baldassarri, M., Furini, Frullanti, E. et al. (2021). Protective Role of a TMPRSS2 Variant on Severe COVID-19 Outcome in Young Males and Elderly Women. *Genes* **12** (4): 596. https://doi.org/10.3390/genes12040596

Mueller, S. N., Rouse, B. T. (2008). Immune Responses to Viruses. *Clinical Immunology*, 421–31. Elsevier. https://doi.org/10.1016/B978-0-323-04404-2.10027-2

Nyholt, D. R. (2004). A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *The American Journal of Human Genetics* **74**(4), pp.765-769. https://doi.org/10.1086/383251

Ovsyannikova, I. G., Haralambieva, I. H., Crooke, S. N., Poland, G. A., Kennedy, R. B. (2020). The Role of Host Genetics in the Immune Response to SARS-CoV-2 and COVID-19 Susceptibility and Severity. *Immunological Reviews* **296** (1): 205–19. https://doi.org/10.1111/imr.12897

Pairo-Castineira, E., Clohisey, S., Klaric, L., Bretherick, A. D., Rawlik, K., Pasko, D., Walker, S., Parkinson, N., Fourman, M. H., Russell, C. D., Furniss, J. (2021). Genetic mechanisms of critical illness in Covid-19. *Nature* **591**(7848), pp.92-98. https://doi.org/10.1038/s41586-020-03065-y

Pendergrass, S. A., Brown-Gentry, K., Dudek, S. M., Torstenson, E. S., Ambite, J. L., Avery, C. L., Buyske, S., Cai, C., Fesinmeyer, M. D., Haiman, C., Heiss, G. (2011). The use of phenome-wide association studies (PheWAS) for exploration of novel genotype-phenotype relationships and pleiotropy discovery. *Genetic epidemiology* **35**(5), pp.410-422. https://doi.org/10.1002/gepi.20589

Pires, D.E., Ascher, D.B. and Blundell, T.L., 2014. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic acids research*, **42**(W1), pp.W314-W319. https://doi.org/10.1093/nar/gku411

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, **38**(8), pp.904-909. https://doi.org/10.1038/ng1847

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/

Ramírez-Bello J, Jiménez-Morales M. (2017). Functional implications of single nucleotide polymorphisms (SNPs) in protein-coding and non-coding RNA genes in multifactorial diseases. *Gac Med Mex* **153**(2):238-250. Spanish. PMID: 28474710

Rasheed, J., Jamil, A., Hameed, A. A., Al-Turjman, F., Rasheed, A. (2021). COVID-19 in the Age of Artificial Intelligence: A Comprehensive Review. *Interdisciplinary Sciences: Computational Life Sciences*, pp.1-23. https://doi.org/10.1007/s12539-021-00431-w

Rescenko, R., Peculis, R., Briviba, M. et al. (2021). Replication of LZTFL1 Gene Region as a Susceptibility Locus for COVID-19 in Latvian Population. *Virol. Sin*. **36**, 1241–1244 https://doi.org/10.1007/s12250-021-00448-x

Santus, E., Marino, N., Cirillo, D., Chersoni, E., Montagud, A., Chadha, A. S., Valencia, A., Hughes, K., Lindvall, C. (2021). Artificial Intelligence–Aided Precision Medicine for COVID-19: Strategic Areas of Research and Development. *Journal of Medical Internet Research* **23**(3), p.e22453. https://doi.org/10.2196/22453

Schoeman, D., Fielding, B. C. (2019). Coronavirus Envelope Protein: Current Knowledge. *Virology Journal* **16** (1): 69. https://doi.org/10.1186/s12985-019-1182-0.

Schroeder, H. W., Cavacini, L. (2010). Structure and Function of Immunoglobulins. *Journal of Allergy and* Clinical *Immunology* **125** (2): S41–52. https://doi.org/10.1016/j.jaci.2009.09.046

Secolin, R., de Araujo, T. K. , Gonsales, M. C., Rocha, C. S., Naslavsky, M., De Marco, L., Bicalho, M. A. C. et al. (2021). Genetic Variability in COVID-19-Related Genes in the Brazilian Population. *Human Genome Variation* **8** (1): 15. https://doi.org/10.1038/s41439-021-00146-w

Setakis, E., Stirnadel, H., Balding, D. J. (2006). Logistic regression protects against population structure in genetic association studies. *Genome research*, *16* (2), pp.290-296. https://doi.org/10.1101%2Fgr.4346306

Severe Covid-19 GWAS Group (2020). Genome-wide association study of severe Covid-19 with respiratory failure. *New England Journal of Medicine* **383** (16), pp.1522-1534. https://doi.org/10.1056/NEJMoa2020283

Shah, V. K., Firmal, P., Alam, A., Ganguly, D., Chattopadhyay, S. (2020). Overview of Immune Response During SARS-CoV-2 Infection: Lessons From the Past. *Frontiers in Immunology* **11** (August): 1949. https://doi.org/10.3389/fimmu.2020.01949

Shelton, J. F., Shastri, A. J., Ye, C., Weldon, C. H., Filshtein-Sonmez, T., Coker, D., Symons, A., Esparza-Gordillo, J., The 23andMe COVID-19 Team, Aslibekyan, S., Auton A. (2021). Trans-ancestry analysis reveals genetic and nongenetic associations with COVID-19 susceptibility and severity. *Nat Genet* **53**(6):801-808. PMID: 33888907. https://doi.org/10.1038/s41588-021-00854-7

Sim, Y., Chung, M. J., Kotter, E., Yune, S., Kim, M., Do, S., Han, K., Kim, H., Yang, S., Lee, D. J., Choi, B. W. (2020). Deep convolutional neural network–based software improves radiologist detection of malignant lung nodules on chest radiographs. *Radiology* **294**(1), pp.199-209. https://doi.org/10.1148/radiol.2019182465

Soremekun, O. S., Omolabi, K. F., Soliman, M. E. (2020). Identification and classification of differentially expressed genes reveal potential molecular signature associated with SARS-CoV-2 infection in lung adenocarcinomal cells. *Informatics in Medicine Unlocked* **20**, p.100384. https://doi.org/10.1016/j.imu.2020.100384

Stoeger, T., Amaral, L. A. N. (2020). Meta-Research: COVID-19 research risks ignoring important host genes due to pre-established research patterns. *Elife* **9**, p.e61981. https://doi.org/10.7554/eLife.61981

Sun, S., Dong, B., Zou, Q. (2021). Revisiting genome-wide association studies from statistical modelling to machine learning. *Briefings in Bioinformatics* **22**(4), p.bbaa263. https://doi.org/10.1093/bib/bbaa263

Suryamohan, K., Diwanji, D., Stawiski, E. W., Gupta, R., Miersch, S., Liu, J., Chen, C., Jiang, Y. P., Fellouse, F. A., Sathirapongsasuti, J. F., Albers, P. K. (2021). Human ACE2 receptor polymorphisms and altered susceptibility to SARS-CoV-2. *Communications biology* **4**(1), pp.1-11. https://doi.org/10.1038/s42003-021-02030-3

Tahamtan, A., Samadizadeh, S., Rastegar, M., Nakstad, B., Salimi, V. (2020). Respiratory syncytial virus infection: why does disease severity vary among individuals? *Expert review of respiratory medicine* **14**(4), pp.415-423. https://doi.org/10.1080/17476348.2020.1724095

Tanimine, N., Takei, D., Tsukiyama, N., Yoshinaka, H., Takemoto, Y., Tanaka, Y., Kobayashi, T., Tanabe, K., Ishikawa, N., Kitahara, Y., Okimoto, M. (2021). Identification of Aggravation-Predicting Gene Polymorphisms in Coronavirus Disease 2019 Patients Using a Candidate Gene Approach Associated With Multiple Phase Pathogenesis: A Study in a Japanese City of 1 Million People. *Critical care explorations* **3**(11). https://doi.org/10.1097/CCE.0000000000000576

Tavasolian, F., Rashidi, M., Hatam, G. R., Jeddi, M., Hosseini, A. Z., Mosawi, S. H., Abdollahi, E., Inman, R. D. (2021). HLA, Immune Response, and Susceptibility to COVID-19. *Frontiers in Immunology* **11** (January): 601886. https://doi.org/10.3389/fimmu.2020.601886

Teuwen, L. A., Geldhof, V., Pasut, A., Carmeliet, P. (2020). COVID-19: the vasculature unleashed. *Nature Reviews Immunology* **20**(7), pp.389-391. https://doi.org/10.1038/s41577-020-0343-0

Toh, C., Brody, J. P. (2020). Evaluation of a genetic risk score for severity of COVID-19 using human chromosomal-scale length variation. *Human genomics* **14**(1), pp.1-5. https://doi.org/10.1186/s40246-020-00288-y

Torre-Fuentes, L., Matías-Guiu, J., Hernández-Lorenzo, L., Montero-Escribano, P., Pytel, V., Porta-Etessam, J., Gómez-Pinedo, U., Matías-Guiu, J. A. (2021). ACE2, TMPRSS2, and Furin variants and SARS-CoV-2 infection in Madrid, Spain. *Journal of medical virology* **93**(2), pp.863-869. https://doi.org/10.1002/jmv.26319

Ulhaq, Z. S., Soraya, G. V. (2020). Anti-IL-6 Receptor Antibody Treatment for Severe COVID-19 and the Potential Implication of IL-6 Gene Polymorphisms in Novel Coronavirus Pneumonia. *Medicina Clínica* **155** (12): 548–56. https://doi.org/10.1016/j.medcli.2020.07.002

V'kovski, P., Kratzel, A., Steiner, S., Stalder, H., Thiel, V. (2021). Coronavirus Biology and Replication: Implications for SARS-CoV-2. *Nature Reviews Microbiology* **19** (3): 155–70. https://doi.org/10.1038/s41579-020-00468-6

Verma, A., Tsao, N., Thomann, L., Ho, Y.L., Carr, R., Crawford, D., Efird, J. T., Huffman, J., Hung, A., Ivey, K., Iyengar, S. (2021). A Phenome-Wide Association Study of genes associated with COVID-19 severity reveals shared genetics with complex diseases in the Million Veteran Program. Preprint. *medRxiv*. https://doi.org/10.1101/2021.05.18.21257396

Visscher, P. M., Brown, M. A., McCarthy, M. I., Yang , J. (2012). Five years of GWAS discovery. *Am J Hum Genet. The American Society of Human Genetics* **90**, pp.7-24. https://doi.org/10.1016/j.ajhg.2011.11.029

Wang, D., Wiktor, S. D., Cheng, C. W., Simmons, K. J., Money, A., Pedicini, L., Carlton, A., Breeze, A. L., McKeown, L. (2022). EFCAB4B (CRACR2A) genetic variants associated with COVID-19 fatality. *medRxiv*. https://doi.org/10.1101/2022.01.17.22269412

Wang, F., Huang, S., Gao, R., Zhou, Y., Lai, C., Li, Z., Xian, W., Qian, X., Li, Z., Huang, Y., Tang, Q. (2020). Initial whole-genome sequencing and analysis of the host genetic contribution to COVID-19 severity and susceptibility. *Cell discovery* **6**(1), pp.1-16. https://doi.org/10.1038/s41421-020-00231-4

Wang ,K., Li, M., Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**(16):e164. https://doi.org/10.1093/nar/gkq603

Wang, R. Y., Guo, T. Q., Li, L. G., Jiao, J. Y., Wang, L. Y. (2020). Predictions of COVID-19 Infection Severity Based on Co-associations between the SNPs of Co-morbid Diseases and COVID-19 through Machine Learning of Genetic Data. *2020 IEEE 8th International Conference on Computer Science and Network Technology (ICCSNT)* (pp. 92-96). IEEE. https://doi.org/10.1109/ICCSNT50940.2020.9304990

Wang, W., Zhang, W., Zhang, J., He, J., Zhu, F. (2020). Distribution of HLA Allele Frequencies in 82 Chinese Individuals with Coronavirus Disease-2019 (COVID-19). *HLA* **96** (2): 194–96. https://doi.org/10.1111/tan.13941.

WEB (a). https://syncedreview.com/2017/10/22/tree-boosting-with-xgboost-why-does-xgboost-win-every-machine-learning-competition/

WHO (2022). World Health Organization. Coronavirus (COVID-19) Dashboard

Wieczorek, M., Abualrous, E. T., Sticht, J., Álvaro-Benito, M., Stolzenberg, S., Noé, F., Freund, C. (2017). Major Histocompatibility Complex (MHC) Class I and MHC Class II Proteins: Conformational Plasticity in Antigen Presentation. *Frontiers in Immunology* **8** (March). https://doi.org/10.3389/fimmu.2017.00292

Williams, F. M., Freydin, M., Mangino, M., Couvreur, S., Visconti, A., Bowyer, R. C., Le Roy, C. I., Falchi, M., Sudre, C., Davies, R., Hammond, C. (2020). Self-reported symptoms of covid-19 including symptoms most predictive of SARS-CoV-2 infection, are heritable. Preprint. *medRxiv*. https://doi.org/10.1101/2020.04.22.20072124

Woolf, B. P. (2010). A roadmap for education technology. *A Report to the Computing Community Consortium:* http://telearn.archives-ouvertes.fr/docs/00/58/82/91/PDF/ groe_roadmap_for_education_technology_final_report_003036v1_.pdf. p. 80 pp.

Wu, L., Zhu, J., Liu, D., Sun, Y., Wu, C. (2021). An integrative multiomics analysis identifies putative causal genes for COVID-19 severity. *Genetics in Medicine*, pp.1-11. https://doi.org/10.1038/s41436-021-01243-5

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6), pp.714-721. https://doi.org/10.1093/bioinformatics/btp041

Wulandari, L., Hamidah, B., Pakpahan, C., Damayanti, N. S., Kurniati, N. D., Adiatmaja, C. O., Wigianita, M. R., Husada, D., Tinduh, D., Prakoeswa, C. R .S., Endaryanto, A. (2021). Initial study on TMPRSS2 p. Val160Met genetic variant in COVID-19 patients. *Human genomics* **15**(1), pp.1-9. https://doi.org/10.1186/s40246-021-00330-7

Wynants, L., Van Calster, B., Bonten, M.M., Collins, G. S., Debray, T. P., De Vos, M., Haller, M. C., Heinze, G., Moons, K. G., Riley, R. D. (2020). Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. *BMJ* **369**: m1328. https://doi.org/10.1136/bmj.m1328

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**(7):565-9. https://doi.org/10.1038/ng.608

Zargari Khuzani, A., Heidari, M., Shariati, S. A. (2021). COVID-Classifier: An automated machine learning model to assist in the diagnosis of COVID-19 infection in chest x-ray images. *Scientific Reports* **11**(1), pp.1-6. https://doi.org/10.1038/s41598-021-88807-2

Zhang, Y., Qin, L., Zhao, Y., Zhang, P., Xu, Li, K., Liang, L. et al. (2020). Interferon-Induced Transmembrane Protein 3 Genetic Variant Rs12252-C Associated With Disease Severity in Coronavirus Disease 2019. *The Journal of Infectious Diseases* **222** (1): 34–37. https://doi.org/10.1093/infdis/jiaa224

Zhao, D., Yao, F., Wang, L., Zheng, L., Gao, Y., Ye, J., Guo, F., Zhao, H., Gao, R. (2020). A comparative study on the clinical features of coronavirus 2019 (COVID-19) pneumonia with other pneumonias. *Clinical Infectious Diseases* **71**(15), pp.756-761. https://doi.org/10.1093/cid/ciaa247

Zhou, W., Nielsen, J. B., Fritsche, L. G. et al. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics* **50,** 1335–1341. https://doi.org/10.1038/s41588-018-0184-y

Zhu, H., Zheng, F., Li, L., Jin, Y., Luo, Y., Li, Z., Zeng, J., Tang, L., Li, Z., Xia, N., Liu, P. (2021). A Chinese host genetic study discovered type I interferons and causality of cholesterol levels and WBC counts on COVID-19 severity. *iScience*. https://doi.org/10.1016/j.isci.2021.103186

Received March 23, 2022, revised September 17, 2022, accepted September 19, 2022

Supplementary tables

Table S1. SARS-CoV-2 entry into the host cell cycle genome variants linked with COVID-19. Abbreviation NA here means “Not Available”.

Gene	Publication	Variant					Sample Size		Laboratory methods
		Protein change	**Accession number***	**Allele Frequency****	**P-value**	**Possible Effect**	**N (cases/controls)**	**Comments**	
ACE2	Suryamohan et al., 2020	Ser19Pro	rs73635825	0.0003	0.0656	Enhanced susceptibility to viral attachment	290,000	400 population groups (gnomAD, RotterdamStudy, ALSPAC, GenomeAsia100k, HGDP, TOMMO-3.5kjpnv2, IndiGen, HGDP databases)	Genotyping
		Ile21Val	rs778030746	0.00001					
		Glu23Lys	rs756231991	0.000005					
		Lys26Arg	rs4646116	0.004					
		Thr27Ala	rs781255386	0.00001					
		Asn64Lys	rs119910071	0.00002					
		Thr92Ile	rs763395248	0.00001					
		Gln102Pro	rs139587809	0.00002					
		His378Arg	rs142984500	0.00009					
		Lys31Arg	rs758278442	0		Decrease attachment propensity to spike protein			
		Asn33Ile	NA	0					
		His34Arg	NA	0					
		Glu35Lys	rs134811469	0.00002					
		Glu37Lys	rs146676783	0.00004					
		Asp38Val	NA	0					
		Tyr50Phe	rs119219261	0.000006					

		Asn51Ser	rs156924369	0.000006					
		Met62Val	rs132554210	0.000006					
		Lys68Glu	rs755691167	0.00001					
		Phe72Val	rs125600725	0.000005					
		Tyr83His	rs759134032	0					
		Gly326Glu	rs759579097	0.000006					
		Gly352Val	rs370610075	0.000006					
		Asp355Asn	rs961360700	0.00001					
		Gln388Leu	rs751572714	0.00002					
		Asp509Tyr	NA	0					
	Benetti et al., 2020	Lys26Arg	rs4646116	0.0039	NA	Impact ACE2 stability	389 (131/258)	Gen-covid multicenter study	Whole exome sequencing
		Gly211Arg	rs148771870	0.0013	NA				
		Leu351Val	NA	0	NA	Interfere with ACE2 and S binding			
		Prp389His	rs762890235	0.000039	NA				
		Val506Ala	rs775181355	0.0000066	NA	Destabilize spike protein and ACE2 interaction			
		Val209Gly	NA	0	NA				
		Gly377Glu	rs767462182	0.0000056	NA				
	Guo et al., 2020	His378Arg	rs142984500	0.0002	NA	Enhanced susceptibility to viral attachment	141,456	Genome Aggregation Database	Whole exome and genome sequencing
		Ser19Pr o	rs73635825	0.003	NA				
		Gly211Ala	rs148771870	0.0012	NA	Affect secondary ACE2 structure			
		Asp206Gly	rs142443432	0.00029	NA				

		Arg219Cys	rs759590772	0.00009	NA				
		Arg219His	rs759590772	0.00009	NA				
		Leu341Arg	rs138390800	0.0004	NA				
		Ile468Val	rs191860450	0.0008	NA				
		Ser547Cys	rs373025684	0.0002	NA				
	Gibson et al., 2020	Lys26Arg	rs4646116	0.00397	NA	Decrease the binding affinity between S protein and ACE2	141,456	Genome Aggregation Database	Whole exome and genome sequencing
		Ser43Arg	rs1447927937	0.000005	NA				
		Gly326Glu	rs759579097	0.000005	NA				
		Met82Ile	rs766996587	0.00001	NA				
		Glu37Lys	rs146676783	0.00003	NA	Increase the binding affinity between S protein and ACE2			
		Thr27Ala	rs781255386	0.00001	NA				
		Lys329Gly	rs143936283	0.00003	NA				
		Lys26Glu	rs1299103394	0.000005	NA				
	Horowitz et al., 2020	NA	rs190509934	0.3	4.5×10^{-13}	Lower risk of COVID-19	756,646 (52,630/704,016)	AncestryDNA COVID-19 Research, Geisinger Health System, Penn MedicineBioBank, UK Biobank	SNP genotyping assay
	MacGowan and Barton, 2020	Gly326Glu	rs759579097	**0.000006**	NA	Enhance ACE2 binding with spike protein	NA	Genome Aggregation Database	Whole exome and genome sequencing
		Glu37Lys	rs146676783	**0.00003**	NA	Weaken ACE2 binding with spike protein			
		Gly352Val	rs370610075	**0.000005**	NA				
		Asp355Asn	**rs961360700**	**0.00001**	NA				

TMPRSS2	Hou Yuan et al., 2020	Val160Met	rs12329760	0.384	0.00005	Susceptibility to COVID-19	81,000	Genome Aggregation Database, Exome Sequencing Project, 1000 Genomes Project	Whole exome and genome sequencing
	Wulandari et al., 2021						95 cases	Patients with moderate and severe COVID-19	SNP genotyping assay
	Monticelli et al., 2021			0.3735	0.0153	Protective effect	1,177 cases	GEN-COVID Multicenter Study	Whole exome sequencing
TLL-1	Grimaudo et al., 2021	NA	rs17047200	**0.13**	0.029	Increased risk of COVID19	383 patients	Mild or severe Sicilian patients	SNP genotyping assay

* Accession number in bold is obtained from https://varsome.com/

** Allele frequency in bold is obtained from https://gnomad.broadinstitute.org/

Table S2. Immune response to SARS-CoV-2 genome variants linked with COVID-19. NA – information not available.

Gene	Publication	Variant					Sample Size		Laboratory methods
		Protein change/ Allele	**Accession number***	**Allele Frequency ****	**P-value**	**Possible Effect**	**N (cases/controls)**	**Comments**	
HLA	Secolin et al., 2021	DRB1*15:01	NA	0.06	NA	Increased	386	BIPMed dataset, ABraOM data set, Brazil	Whole exome sequencing
		DQB1*06:02	NA	0.08	NA				
		B*27:07	NA	0.001	NA				
	Wang F. et al., 2020	C*07:29	NA	**0.0001**	1.00×10^{-3}		3,872 (82/3,790)	The same study	Next-generation sequencing
		B*15:27	NA	**0.00004**	1.00×10^{-3}				
IFITM3	Zhang et al., 2020	Ser14=	rs12252	**0.13**	9.30×10^{-3}		80 cases	Beijing Youan Hospital, Capital Medical University, Beijing	*IFITM3* sequencing
PNPLA3	Grimaudo et al., 2021	Ile148Met	rs738409	**0.28**	3.50×10^{-2}		383 patients	Mild or severe Sicilian patients	SNP genotyping assay
APOE	Kuo et al., 2020	Cys130Arg	rs429358	**0.22**	NA		322,948 (622/323,570	UK Biobank	SNP genotyping assay
		Arg176Cys	rs7412	**0.101**	NA				
ABO	Severe Covid-19 GWAS Group, 2020	NA	rs657152	**0.44**	5.00×10^{-8}		3,815 (1,610/2,205)	Seven centers in the Italian and Spanish epicenters	SNP genotyping assay
LZTFL1		NA	rs11385942	**0.08**	1.00×10^{-10}				
	Rescenko et al., 2021	NA	rs71325088	**0.05**	7.00×10^{-3}		2,692 (475/2,217)	Genome Database of Latvian Population	SNP genotyping assay
		NA	rs11385942	**0.068**	5.00×10^{-5}				
		NA	rs73064425	**0.056**	7.00×10^{-3}				

Gene	Publication	Variant					Sample Size		Laboratory methods
		Protein change/ Allele	Accession number*	Allele Frequency **	P-value	Possible Effect	N (cases/controls)	Comments	
SLC6A20	The COVID-19 Host Genetics Initiative, 2021	NA	rs2271616	0.118	1.79×10^{-34}		2,049,562 (49,562/2,000,000)	COVID-19 Host Genetics Initiative	SNP genotyping assay
ABO		NA	rs912805253	0.65	1.45×10^{-39}				
RPL24		NA	rs11919389	0.352	3.46×10^{-15}				
PLEKHA4		NA	rs4801778	0.18	1.18×10^{-8}				
LZTFL1		NA	rs10490770	0.085	9.72×10^{-30}				
FOXP4		NA	rs1886814	0.047	2.41×10^{-8}				
TMEM65		NA	rs72711165	0.013	2.13×10^{-9}				
OAS1		NA	rs10774671	0.66	1.61×10^{-11}				
KANSL1		NA	rs1819040	0.18	1.83×10^{-10}				
TAC4		NA	rs77534576	0.033	4.37×10^{-9}				
DPP9		NA	rs2109069	0.31	4.08×10^{-9}				
RAVER1		NA	rs74956615	0.04	1.94×10^{-4}				
IFNAR2		NA	rs13050728	0.65	2.72×10^{-20}				
LZTFL1	Pairo-Castineira et al., 2021	NA	rs73064425	**0.0761**	4.77×10^{-30}		10,056 (1,676/8,380)	European descent from GenOMICC, UK Biobank	SNP genotyping assay
CCHCR1		NA	rs143334143	**0.14**	8.82×10^{-18}				

Gene	Publication	Variant					Sample Size		Laboratory methods
		Protein change/ Allele	**Accession number***	**Allele Frequency ****	**P-value**	**Possible Effect**	**N (cases/controls)**	**Comments**	
OAS3		NA	rs10735079	**0.75**	1.65×10^{-8}				
RAVER1		NA	rs74956615	**0.04**	2.31×10^{-8}				
IFNAR2		NA	rs2236757	**0.77**	5.00×10^{-8}				
SLC6A20	Shelton et al., 2021	NA	rs2531743	**0.84**	7.60×10^{-10}		114,240 (12,972/101,268)	23andMe	SNP genotyping assay
		NA	rs13078854	0.867	1.60×10^{-18}				
ABO		NA	rs9411378	**0.22**	5.30×10^{-20}				
SLC6A20	Ma et al., 2021	NA	rs11385942	**0.07**	2.87×10^{-16}		680,128 (3,288/676,840)	GWAS summary data from Ellinghaus et al. and COVID-19 Host Genetic Consortium	SNP genotyping assay
ABO		NA	rs8176719	**0.38**	4.00×10^{-7}				
		NA	rs657152	**0.43**	5.53×10^{-6}				
IFNAR2-IL10RB		NA	rs9976829	**0.77**	2.57×10^{-6}				
DNAH /SL C39A10	Hu et al., 2021	NA	rs73060484	0.069	6.00×10^{-4}		1,096 (292/804)	UK Biobank	Genotyping
		NA	rs77578623	0.070	6.20×10^{-4}				
		NA	rs74417002	0.034	3.00×10^{-2}				
		NA	rs73070529	0.048	3.60×10^{-4}				
		NA	rs113892140	0.044	2.80×10^{-3}				

Gene	Publication	Variant					Sample Size		Laboratory methods
		Protein change/ Allele	Accession number*	Allele Frequency **	P-value	Possible Effect	N (cases/controls)	Comments	
		NA	rs200008298	0.032	3.10×10^{-2}				
		NA	rs183712207	0.007	7.70×10^{-3}				
		NA	rs191631470	0.007	3.90×10^{-2}				
CLUAP1		NA	rs2301762	0.055	2.00×10^{-5}				
DES/SPEG		NA	rs71040457	0.355	7.70×10^{-3}				
STXBP5		NA	rs117928001	0.049	1.10×10^{-5}				
		NA	rs116898161	0.046	6.90×10^{-5}				
TOMM7		NA	rs13227460	0.278	2.60×10^{-2}				
		NA	rs55986907	0.286	3.50×10^{-5}				
WSB1		NA	rs60811869	0.024	6.50×10^{-4}				
		NA	rs117217714	0.013	3.30×10^{-5}				
PCDH15		NA	rs9804218	0.357	3.30×10^{-3}				
CPQ		NA	rs7817272	0.194	1.70×10^{-5}				
		NA	rs4735444	0.201	5.80×10^{-6}				
		NA	rs1431889	0.193	3.50×10^{-5}				

Gene	Publication	Variant					Sample Size		Laboratory methods
		Protein change/ Allele	**Accession number***	**Allele Frequency ** **	**P-value**	**Possible Effect**	**N (cases/controls)**	**Comments**	
		NA	rs2874140	0.194	4.00×10^{-5}				
		NA	rs531453964	0.185	3.20×10^{-6}				
		NA	rs7007951	0.184	4.40×10^{-5}				
		NA	rs920576	0.201	1.60×10^{-4}				
OAS1-3	Pairo-Castineira et al., 2021	NA	rs10735079	0.755	1.65×10^{-8}		100,000+ (2,244, 100,000+)	GenOMICC and ISARIC 4C studies	Genotyping, whole genome sequencing
TYK2		NA	rs74956615	0.047	2.30×10^{-8}				
DPP9		NA	rs2109069	0.300	3.98×10^{-12}				
IFNAR2		NA	rs2236757	0.770	4.99×10^{-8}				
ABO	Verma et al., 2021	NA	rs550057	0.240	NA		455,683	EHR and genomic data from two biobanks: Veteran Affairs Million Veteran Program (VAMVP), United Kingdom Biobank (UKBB)	Genotyping
		NA	rs505922	0.350	NA				
RAVER1		NA	rs74956615	0.047	NA				
TLR7	Fallerini et al., 2021	Ser301Pro	NA	NA	NA		156 (79, 77)	From Italian GEN-COVID (Daga et al., 2021)	Genotyping
		Arg920Lys	rs189681811	0.0002	NA				
		Ala1032Thr	rs147244662	0.0006	NA				
		Val219Ile	rs149314023	0.0003	NA				

Gene	Publication	Variant					Sample Size		Laboratory methods
		Protein change/ Allele	Accession number*	Allele Frequency **	P-value	Possible Effect	N (cases/controls)	Comments	
		Ala288Val	rs200146658	0.000012	NA				
		Ala448Val	rs5743781	0.00465	NA				
OAS-1	Tanimine et al., 2021	NA	rs1131454	0.41	0.0048		230 cases	3 Hospitals in Hiroshima, Japan	Genotyping
IL1B		NA	rs1143627	0.48	0.0207				
EFCAB4B	Wang D. et al., 2022	Ala98Thr	rs17836273	0.123	0.012		500,000 (10,118/489,882)	UK Biobank	Genotyping
		His212Gln	rs36030417	0.118	0.013				
		Arg7Gly	rs9788233	0.1453	0.004				
NLRP3	Maes et al., 2022	NA	rs10157379	**0.6067**	0.106		528 cases	University Hospital of Londrina (HU) and the Emergency Rooms (ER) in Londrina, Paraná, Brazil	Genotyping
		NA	rs10754558	**0.6325**	0.167				
ADAM9	Carapito et al., 2021	NA	rs7840270	**0.434**	0.017		72 (47/25)	University hospital network in northeast France (Alsace)	Gene expression
OAS1	Huffman et al., 2021	NA	rs10774671	**0.672**	0.03	Decreased	120,473 (1,842/118,631)	COVID-19 patients of African ancestry	Genotyping

* Accession number in bold is obtained from https://varsome.com/

** Allele frequency in bold is obtained from https://gnomad.broadinstitute.org/ or http://www.allelefrequencies.net/