UNIVERSITÀ DI SIENA 1240

Department of Medical Biotechnologies

Doctorate in Genetics, Oncology and Clinical Medicine (GenOMeC)

XXXVI Cycle (2020-2023)

Coordinator: Prof. Ilaria Meloni

Host genetic basis of COVID-19: from phenotype to genes.

Scientific disciplinary sector: MED/03 - Medical Genetics

Candidata

Margherita Baldassarri

Firmato digitalmente da
Margherita Baldassarri
Data: 11.12.2023 15:14:52 CET

Tutor

Prof. Francesca Mari

Supervisor

Prof. Simone Furini

Index

1. Abstract

COVID-19 pandemic represented a worldwide challenge in the past 3 years. Even if the emergency is ending, the identification of the main prognostic factors that determine the clinical outcome still remain one of the major challenges, to implement effective preventive and therapeutic strategies also useful for other infectious diseases and complex diseases. Retrospective analysis of epidemiologic data has revealed that male sex, older age, and underlying metabolic diseases such as obesity pose high risks for fatal COVID-19. It is now well known that also host genetics factors play a crucial role in causing the COVID-19 clinical outcome and various common and rare variants have been described in association with susceptibility and severity.

In particular, within the GEN-COVID Multicenter study, from March 2020, we have enrolled more than 5000 Sars-CoV-2 positive subjects, collected their clinical data and performed WES analysis. Through data-mining from clinical and molecular data, during these three years of my PhD course, we have discovered that common, rare and ultra-rare variants in genes such as *TLR7, CFTR, AR, TLR3*, *SELP* and *ADAMTS13* are associated with the severity of COVID-19 disease through different mechanisms.

Moreover, in order to better understand the genetic basis of COVID-19, as a complex, polygenic disease, it was therefore necessary to apply new approaches capable of identifying the entire genetic variability of the host and combining common and rare variants in a single model. For this reason, we proposed a new model, the post-Mendelian model, for a genetic characterization of the disorder based on an adapted Poly-genic Risk Score (PRS), called Integrated PolyGenic Score (IPGS). This allowed us to reach a more precise disease severity prediction than that based on sex and age alone.

Modelling precisely the role of the entire range of host genomics affecting disease susceptibility and severity in COVID-19 is critical to obtaining a complete biological understanding of the aetiology and pathogenicity of COVID-19 as well as other severe complex diseases The knowledge that will be acquired with our data will allow us to develop a new therapeutic approach, which acts through the host-mediated response to pathogens rather than acting directly on the pathogen, as standard therapy traditionally addresses.

2. Introduction

2.1 SARS-Cov-2

Since 2002, three major outbreaks of coronavirus, a zoonotic virus known to cause respiratory illness, have been reported, including SARS-CoV (Severe Acute Respiratory Syndrome Coronavirus 1), MERS-CoV (Middle East respiratory syndrome), and the most recent 2019-nCoV, more commonly known as SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2).

The first cases of this novel coronavirus SARS-CoV-2 were detected, in December 2019, in Wuhan, Hubei province, central China. Then the virus spread rapidly to other countries across the world. This led World Health Organization (WHO) to declare a Public Health Emergency of International Concern (PHEIC) on 30 January 2020, and to characterize the outbreak as a pandemic on 11 March 2020. The disease caused by SAR-CoV-2 was termed Coronavirus Disease-2019 (COVID-19) (https://www.who.int, The New York Times, 2020; Wang C. et al., 2020). This disease, characterized by a high transmissibility, has quickly turned into a global health emergency.

Based on the genome sequence analysis, the SARS-CoV-2 strain being classified under the beta-coronavirus genus (β-CoV) that together with alpha-coronavirus (α-CoV), delta-coronavirus (δ-CoV) and gamma-coronavirus (γ-CoV) belong to the Coronavirinae subfamily. Among them, α- and β- CoVs infect mammals, γ-CoVs infect avian species, and δ-CoVs infect both mammals and avian species. (Wang L. et al., 2014; Zhu N. et al., 2019).

α-CoVs and β-CoVs are the only ones among the CoVs of zoonotic origin that are able to infect humans, for which they are highly pathogenic (Cui J. et al., 2019; Luk H.K. et al., 2019). The two highly pathogenic β-CoVs, SARS-CoV and MERS-CoV, cause severe respiratory syndrome in humans, and the other four human coronaviruses (HCoV-NL63, HCoV-229E, HCoV-OC43 and HKU1) induce only mild upper respiratory diseases in immunocompetent hosts, although some of them can cause severe infections in infants, young children and elderly individuals (Su S. et al., 2016; Forni D. et al., 2017).

The SARS-CoV epidemic spread in 2002 in China with a mortality rate of 11%, while the MERS-CoV epidemic first emerged in Saudi Arabia and later spread to other countries with a mortality rate of 37% (Song Z. et al., 2019; Graham R.L. et al., 2013; Zumla A. et al., 2015; Hui D.S. et al., 2018; Su S. et al., 2015).

The CoV genome is known to have a 5′ cap and a 3′ poly (A) tail; therefore, upon infecting the host cell, the genome acts as an mRNA for translation of the replicase polyproteins required for viral replication (Sawicki S.G. et al., 2007).

CoVs are positive-stranded RNA viruses, with the largest viral genome of the RNA viruses (27 to 33 kb). The single-stranded genome is capped and poly-adenylated. The virus particle is enveloped and carries extended spike proteins on the membrane surface, providing the typical crown-like structure (crown=corona) seen by electron microscopy (Pyrc K. et al., 2007).

Analysing the genome sequence, The SARS-CoV-2 genome share about 82% sequence identity with SARS-CoV and MERS-CoV and >90% sequence identity for essential enzymes and structural proteins. The SARS-CoV-2 genome includes 14 open reading frames (ORFs), of which two-thirds encode for 16 non-structural proteins (nsp 1-16) that constitute the replicase complex. The remaining ORFs encode for 9 accessories and 4 structural proteins: Spike (S), envelope (E), membrane (M), and nucleocapsid (N). These gene products play important roles in viral entry, fusion, and survival in host cells (Smith E.C. et al., 2013; Wang L. et al., 2014; Zhu N et al., 2019; Zhang T. et al., 2020; Weiss S.R. et al., 2005, Petrovszki D. et al., 2022) (**Figure 1**).

Notably, the S protein is the one that mediates virus entry into host cells. The S protein binds to the host receptor through the receptor-binding domain (RBD) in the S1 subunit, followed by the fusion of the S2 subunit to the cell membrane. Different cell surface receptors recognize RBD of S proteins of SARS-CoV and MERS-CoV. MERS-CoV recognizes the dipeptidyl peptidase 4 receptor. Whereas, SARS-CoV and SARS-CoV-2 recognize the ACE2 receptor to bind with the viral S protein. These CoVs mainly differ in their mechanism of host entry, suggesting possible changes in the residual composition of S protein that may dictate host entry (Naqvi A.A.T. et al., 2020) (**Figure 1**).

The S1 subunit is involved in the attachment of virions with the host cell membrane by interacting with human ACE2 that subsequently initiates the infection process. During this process, S protein undergoes conformational changes induced upon its entry into the endosomes of the host cell. Mutations in the S protein seem to induce conformational changes, which may cause an altered antigenicity. Although several mutations have been found in the S1 receptor binding region of SARS-CoV-2, its interaction with ACE2 is preserved in humans, swine, civet, and bats, except for mouse ACE2.

The S2 subunit works as the fusion protein that helps in the fusion of virion with the mammalian cell membrane. During the process of fusion, the S2 protein appears in three

main conformational states 1) pre-fusion native state, followed by 2) a pre-hairpin intermediate state, and 3) ensuing post-fusion hairpin state. The remaining S2' cleaved subunit of the S protein functions as a fusion peptide (Naqvi A.A.T. et al., 2020) (**Figure 1**).

High ACE2 expression was identified in type II alveolar cells of lung, kidney proximal tubule cells, and bladder urothelial cells, oesophagus upper and stratified epithelial cells, absorptive enterocytes from ileum and colon, cholangiocytes, and myocardial cells. These findings indicated that those organs with high ACE2-expressing cells should be considered as potential high risk for infection (Xu H. et al., 2020).

While the S1 subunit of the spike protein is responsible for anchoring the virion by binding to the ACE2 cellular receptor of the host cell, the S2 subunit enhances the fusion of the viral and the host cell membranes. The fusion is mediated by the S2 subunit that is activated by the transmembrane protease serine 2 (TMPRSS2) cleaving the spike protein at the S1/S2 sites (Hoffmann M. et al., 2020, Petrovszki D. et al., 2022), (**Figure 1**).

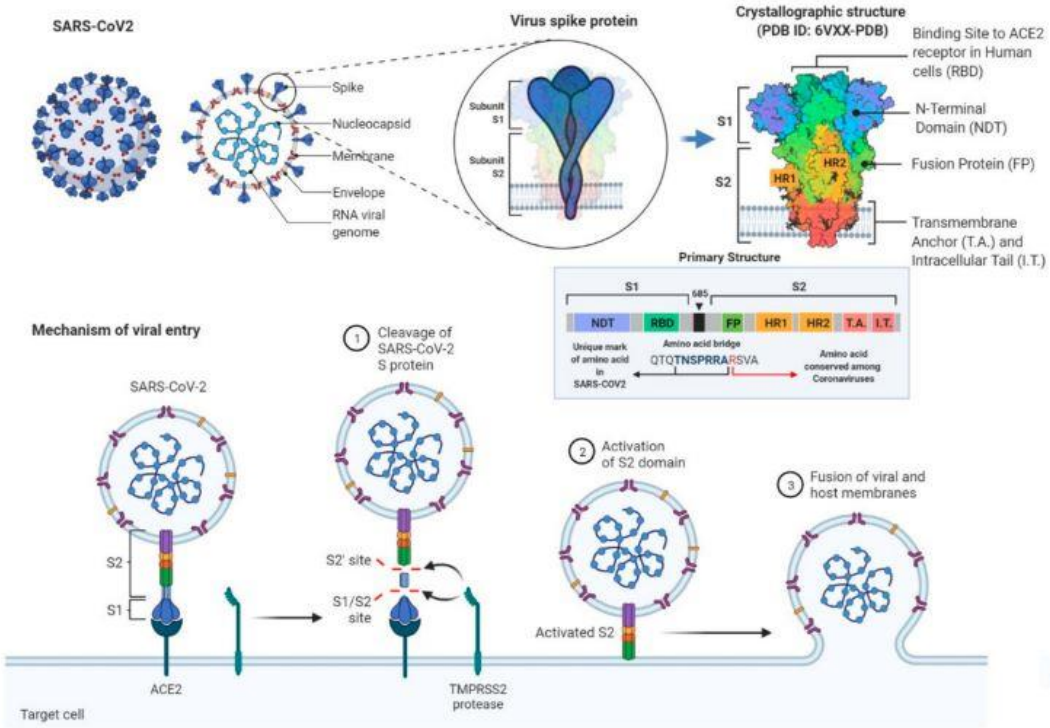


Figure 1. **SARS-CoV-2 structure and mechanism of viral entry.** Spike protein plays a crucial role in this process. While S1 subunit is responsible for anchoring the virion by binding to the cellular receptor angiotensin-converting enzyme 2 (ACE2) of the host cell, S2 subunit enhances the fusion of the viral and the host cell membranes. The fusion is mediated by the S2 subunit that is activated by the transmembrane protease serine 2 (TMPRSS2) cleaving the spike protein at the S1/S2 sites. From Petrovszki D. et al., 2022

TMPRSS2 facilitates entry of SARS-CoV-2 into the cell, while the Furin acts as a Spike activator in endosomes, thereby rendering even cells lacking TMPRSS2 susceptible.

Once the virus has established contact with the host cell, the genome is released and viral replicase proteins are produced; these elements reorganize the cellular endoplasmic reticulum into double-membrane vesicles (DMVs) that assist in the replication of viral genomic and sub-genomic RNAs (sgRNAs) required for viral particle formation (Ashraf U.M. et al., 2021).

E proteins are a group of relatively small viral proteins that help in the assembly and release of the virions. The E protein is relatively small (75 aa), and plays a significant role in the viral morphogenesis and assembly. Acts as viroporins that assemble into host membrane forming protein-lipid pores involved in ion transport.

M proteins are 222 amino acid long structural proteins that function in concurrence with E, N, and S proteins, and plays a major role in the RNA packaging.

N protein play an important role in the packaging of viral RNA into ribonucleocapsid. It mediates viral assembly by interacting with the viral genome and M protein, which are helpful in the augmentation of viral RNA transcription and replication (Naqvi A.A.T. et al., 2020).

2.2 COVID-19: Transmission, pathogenesis and clinical features

COVID-19 transmission has been identified to originate from bats but may have been transmitted to humans through other intermediate animals potentially sourced from the local seafood market in Wuhan city, Hubei province, China. Xiao et al. stated that for SARS-CoV-2 to transmit to humans, an intermediate host must always be present, as bat-derived CoVs rarely infect humans. The study also reported that the majority of Chinese and Malayan wild pangolins, tested for SARS-CoV-2-like coronaviruses, have been positive. It is important to highlight that after thorough analysis, a single receptor-binding domain (RBD) in the spike protein of the Pangolin-CoV was found to have a minor difference in only one amino acid from that of SARS-CoV-2. These data further indicated that the SARS-CoV-2 may have potentially originated from the viral recombination between Pangolin-CoV and Bat-nCoV before transmitting to humans (Xiao K. et al., 2020), (**Figure 2**).

It is now accepted that the main form of human-to-human transmission happens through respiratory droplets expelled by an infected individual; therefore, coughing and sneezing renders SARS-CoV-2 airborne, putting non-infected individuals at risk of contracting the disease (Carlos W. et al., 2020; Ather A. et al., 2020). More recently, it has been highlighted that SARS-CoV-2 has been detected in the fecal samples of infected

patients, indicating the ability of SARS-CoV-2 to proliferate within the digestive tract and the potential for fecal–oral route of transmission. Moreover, it has been reported that SARS-CoV-2 transmission can also occur as a result of contact with contaminated inanimate objects, also known as fomite transmission. Indeed, on plastic or stainless-steel surfaces, they can remain stable for up to days.

Hospitals are known to be one of the sources of secondary SARS-CoV-2 transmission, as they host a large number of infected individuals (Drosten C. et al., 2014). In the case of COVID-19, viral contamination in hospital rooms where COVID-19 patients are being cared for has been reported as another mode of transmission (Sharma A. et al., 2021 - **Figure 3**).

One of the routes of transmission that has had the greatest impact on the spread of SARS-CoV-2 and makes it so difficult to track is through asymptomatic people; it appears that infected people who have no symptoms, even the mildest, are able to transmit the virus over a period of several days, making case tracking difficult if not impossible (Wang C et al., 2020). On infection, the median incubation period is approximately 4–5 days before symptom onset, with 97.5% of symptomatic patients developing symptoms within 11.5 days (Tay M.Z. et al., 2020).

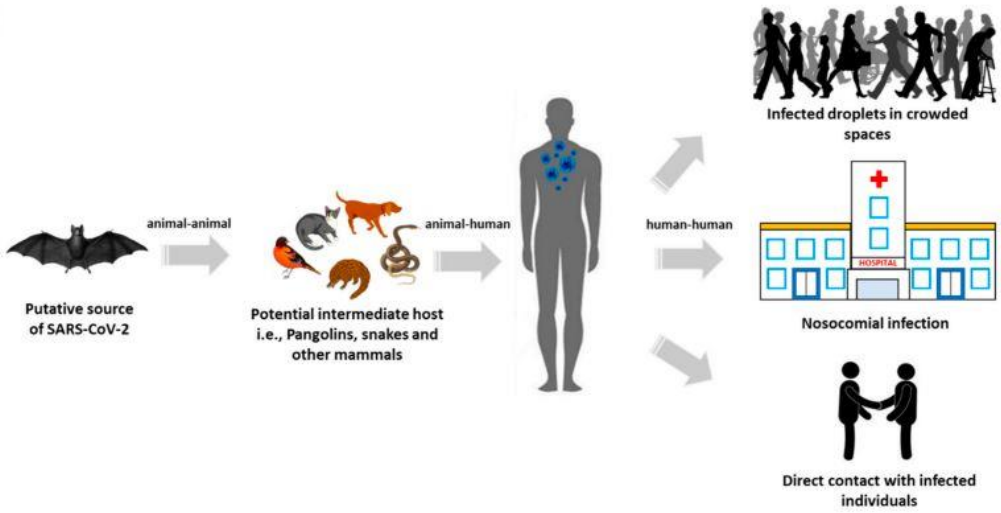


Figure 3. **Representation of the zoonotic COVID-19 modes of transmission.** Modified from Sharma A. et al., 2021.

COVID-19 disease can present very differently from person to person, in a spectrum ranging from asymptomatic cases, in which the infection is detected by PCR or antigen testing but produces no symptomatic effect, to cases in which the disease rapidly progresses to acute respiratory distress syndrome (ARDS) and the patient requires mechanical

ventilatory support and even dies. Within this very wide range there are also intermediate cases, in which the infection presents with a multitude of non-specific symptoms that may necessitate therapeutic control.

Although the symptoms of COVID-19 disease are often described as predominantly respiratory, as already reported above, ACE2 is not exclusively expressed in the lungs, but is present in most tissues of the body, such as the small intestine, kidneys, heart, thyroid, testis, and adipose tissue, indicating the virus may directly infect cells of other organ systems when viremia is present This means that COVID-19 symptoms are far from being restricted to the respiratory mucosa, but instead affect various body districts and systems.

The most common respiratory symptoms include cold, dry, wheezing, non-productive cough, sore throat, dyspnoea and breathlessness, pneumonia; in addition to these, the most common symptoms also include fever, malaise, headache, and asthenia. Moreover, many patients present with gastrointestinal symptoms (such as diarrhoea, vomiting) or neurological symptoms (such as ageusia, anosmia).

ACE2 is crucial for maintaining tissue homeostasis and negatively regulates the renin-angiotensin-aldosterone system (RAAS) in humans. The RAAS is paramount for normal function in multiple organ systems including the lungs, heart, kidney, and vasculature. Given that SARS-CoV-2 internalizes via ACE2, the resultant disruption in ACE2 expression can lead to altered tissue function and exacerbate chronic diseases, influencing blood pressure and fluid/electrolyte balance, and enhancing inflammation and vascular permeability in the airways.

In order to understand the complexity of COVID-19 disease, it is necessary to assess the effects of the virus on different body tissues and areas (**Figure 4**):

- *LUNGS*: SARS-CoV-2 infects well-differentiated cells on the apical surface of the lungs through binding to ACE2. Thus, SARS-CoV-2 results in susceptibility to the acute respiratory distress syndrome (ARDS). This syndrome evolves in three main phases, which are driven by different pathogenic mechanisms. The so-called exudative phase (phase I) is caused by the initial response of the lungs to injury, which then present an inflammation of the alveolar epithelium. This damage is then managed by the innate cell-mediated immunity of the alveolar endothelium and epithelial barriers, resulting in the accumulation of protein-rich oedematous fluid and pulmonary fluid clearance in the interstices and alveoli. As a result, there is an imbalance in surfactant function, which causes increased surface tension in the lungs,

leading to alveoli collapse and hypoxia. At this point, the proliferative phase of ARDS (phase II) occurs, which involves the repair process; through the expansion of fibroblasts, tissue homeostasis is regenerated. Finally comes the fibrotic phase (phase III), in which we have damage to the cell basement membrane and late re-epithelialisation of the lung tissue, leading to the formation of intra-alveolar fibrotic tissue. This phase, which does not occur in all patients, is linked to prolonged mechanical ventilation and increased mortality (Ashraf U.M. et al., 2021). Patients who develop ARDS are generally older and often have multiple comorbidities, such as hypertension, chronic respiratory disease, cardiovascular disease, diabetes or neoplasia. It should be noted that many of these comorbidities are characterised, at the molecular level, by shifts in the ACE/ACE2 regulatory balance.

- *ORAL CAVITY AND TONGUE*: The oral cavity is not associated with severe symptoms in COVID-19 disease, however ACE2 is diffusely expressed in the oral mucosa and is enriched in oral epithelial cells (Ashraf U.M. et al., 2021). Two of the most characteristic symptoms of COVID-19 are ageusia (loss of taste) and anosmia (loss of smell). Of note, the olfactory and gustatory dysfunctions are common symptoms at the beginning of the infection and gradually disappear.

- *ENDOTHELIUM*: ACE2 is widely expressed on vascular endothelial cells; it is responsible for the conversion of angiotensin II to angiotensin (1-7), which has the effect of activating endothelial nitric oxide synthase (eNOS) in the veins. This enzyme is responsible for the production of nitric oxide, a metabolite with an essential role in vasodilation (Ashraf U.M. et al., 2021). Both ACE2 and angiotensin (1-7) play a protective role in the vascular system, protecting the endothelium and inhibiting the inflammatory response; ACE2 deficiency results in an imbalance in nitric oxide production, which decreases the bioavailability of nitric oxide used by smooth muscle cells to promote vasodilation. Once the virus is able to replicate in cells, the expression of ACE2 is downregulated, consequently decreasing its vasodilatory effect in the vasculature. Prolonged contraction of the arteries is accompanied by their malfunctioning and increased inflammation, thus leading to non-negligible vascular damage (Ashraf U.M. et al., 2021).

- *HEART*: The mechanism of acute myocardial injury caused by SARS-CoV-2 infection might be related to ACE2. ACE2 is widely expressed also in the cardiovascular system and, therefore, ACE2- related signalling pathways might also have a role in heart injury. Other proposed mechanisms of myocardial injury include a cytokine storm triggered by an imbalanced response by type 1 and type 2 T helper cells, and respiratory dysfunction and hypoxemia. Patients with COVID-19 disease and myocardial damage showed oedema of the myocardial stroma, inflammatory cell infiltration and atrophy of cardiac muscle fibres (Zheng Y et al., 2020). The ACE2/angiotensin-(1-7)/MAS axis plays a beneficial role for the heart by inducing vasodilation of the coronary veins, participating in the inhibition of oxidative stress, aiding in the recovery of post-ischemic cardiac function and attenuating pathological cardiac remodelling. Indeed, *ACE2* expression is increased in the early stages of cardiac damage but tends to decrease as the damage progresses. Thus, SARS-CoV-2-induced down-expression of ACE2 has a negative effect on protection against cardiac damage. Furthermore, over-regulation of Ang-II associated with downregulation of *ACE2* in COVID-19 patients results in hyper-activation of RAAS and loss of protective effects of angiotensin-(1-7), worsening cardiac damage (Zheng Y at al., 2020).

- *KIDNEYS*: kidneys play a key role in maintaining fluid and electrolyte homeostasis and controlling blood pressure. ACE2 in the kidneys is expressed at the epithelial level, and its loss leads to an increase in sodium reabsorption and an increase in blood volume, which in turn leads to an increase in blood pressure and possible kidney damage. Susceptibility to SARS-CoV-2 occurs because an alternative pathway of RAAS involves angiotensin (1-7) being produced by ACE2, with an opposite effect to that of angiotensin II, thus leading to a decrease in blood pressure by excretion of sodium. The internalisation of ACE2 caused by SARS-CoV-2 causes an imbalance in the ratio of angiotensin II to angiotensin (1-7), emphasising possible renal and carotid damage, and consequently increasing mortality. Renal involvement is frequently observed in COVID-19, varying from mild proteinuria and minor serum creatinine elevations to acute kidney injury and renal failure. Direct SARS-CoV-2 infection of the renal epithelium is estimated to result in mitochondrial dysfunction, acute tubular necrosis, and protein leakage. In addition to direct infection, uncontrolled cytokine release, thrombosis, and ischemia can also result in further

kidney dysfunction, characterized by intra-renal inflammation, increased vascular permeability, and volume depletion (Bohn M.K. et al., 2020).

- *GASTROINTESTINAL TRACT and LIVER:* The involvement of the gastrointestinal tract and hepatic system in COVID-19 disease progression is being increasingly reported. The most common manifestations reported COVID-19 patients include diarrhoea, nausea, vomiting, and abdominal pain. There is evidence of direct SARS-CoV-2 GI infection through isolation of viral RNA from gastrointestinal epithelial cells. It is thus hypothesized that these manifestations are a result of SARS-CoV-2 infection of intestinal enterocytes and subsequent dysfunction in the ileum and colon. Several studies have reported elevated liver enzymes and higher rates of liver injury in patients with severe COVID-19. Patients with abnormal liver function tests, particularly elevated alanine aminotransferase (ALT) and aspartate aminotransferase (AST), also had significantly higher risk of developing severe pneumonia. Several potential pathophysiological mechanisms have been proposed: first, there is potential for ACE2-mediated liver dysfunction. Although hepatocytes have not been shown to exhibit high ACE2 expression, a high level of ACE2 expression has been reported in cholangiocytes, suggesting direct bile duct infection/damage as a potential cause of abnormal liver enzymes. Otherwise, contrary to earlier studies, more recently it has been observed abundant SARS-CoV-2 viral particles in hepatocytes of post-mortem specimens. Additional pathophysiological mechanisms underlying liver injury include drug-induced liver injury as well as hypoxic hepatitis, and systemic inflammatory response, leading to cytotoxic T-cell-mediated necrosis and MOF. Moreover, the pleiotropic hepatic effects of IL-6 could play a particularly important role, inducing expression of serum amyloid A, fibrinogen, and CRP (Bohn M.K. et al., 2020).

- *PANCREAS*: Pancreatic injury has also been reported in patients with COVID-19. Single-cell RNA sequencing suggests that ACE2 is expressed in both the exocrine and islet cells of the pancreas. In terms of exocrine-related damage, patients could have serologic evidence of exocrine-pancreatic injury, defined as elevated amylase or lipase. Hyper-lipasemia was also reported in COVID-19 patients. Acute diabetes has been observed in SARS-CoV-2 patients. Although direct damage of pancreatic β-cells has been proposed as a plausible mechanism behind this phenotype, immune

destruction of β-cells has also been suggested in addition to bystander death due to exocrine infection. Importantly, COVID-19 appears to enhance complications in patients with diabetes, likely due to viral-induced pancreatic dysfunction as well as associated immune dysregulation, vasculopathy, and coagulopathy (Bohn M.K. et al., 2020).

- *BRAIN*: The reported neurological manifestations of COVID-19 include headache, dizziness, confusion, epilepsy, ataxia (lack of voluntary muscle movement), hyposmia/anosmia, ageusia, and Guillain-Barré syndrome, among others. Altered sense of taste or smell can be present in up to 80% of patients with mild to moderate COVID-19. The underlying pathophysiology of the loss of these olfactory and gustatory perceptions have been postulated to be related to direct damage of the supporting cells of the olfactory epithelium, olfactory bulb and altered function of the olfactory neurons, altered ACE2 signal transmission, and accelerated gustatory particle degradation by sialic acid. It was possible to analyse the effect of SARS-CoV-2 on brain structure by autopsy of patients who died of COVID-19 disease. Autopsy findings in SARS-CoV infections have shown strong evidence of neuro-invasion, with demonstrated viral presence in the cerebro-spinal fluid. Potential mechanisms include: 1) viral entry via ACE2 receptors into the endothelia that line the blood capillaries and subsequent neuro-invasion, 2) neurological oedema and brain stem compression as a result of breached blood-brain barrier,3) neuro-logical oedema and hypercoagulability as a result of cytokine storm syndrome, and 4) propagation via mechanoreceptors and chemoreceptors in the lung and lower respiratory airways. Importantly, it is possible that the neurological manifestations of COVID-19 could be a result of hypoxia, respiratory, and/or metabolic acidosis at end-stage disease (Bohn M.K. et al., 2020).

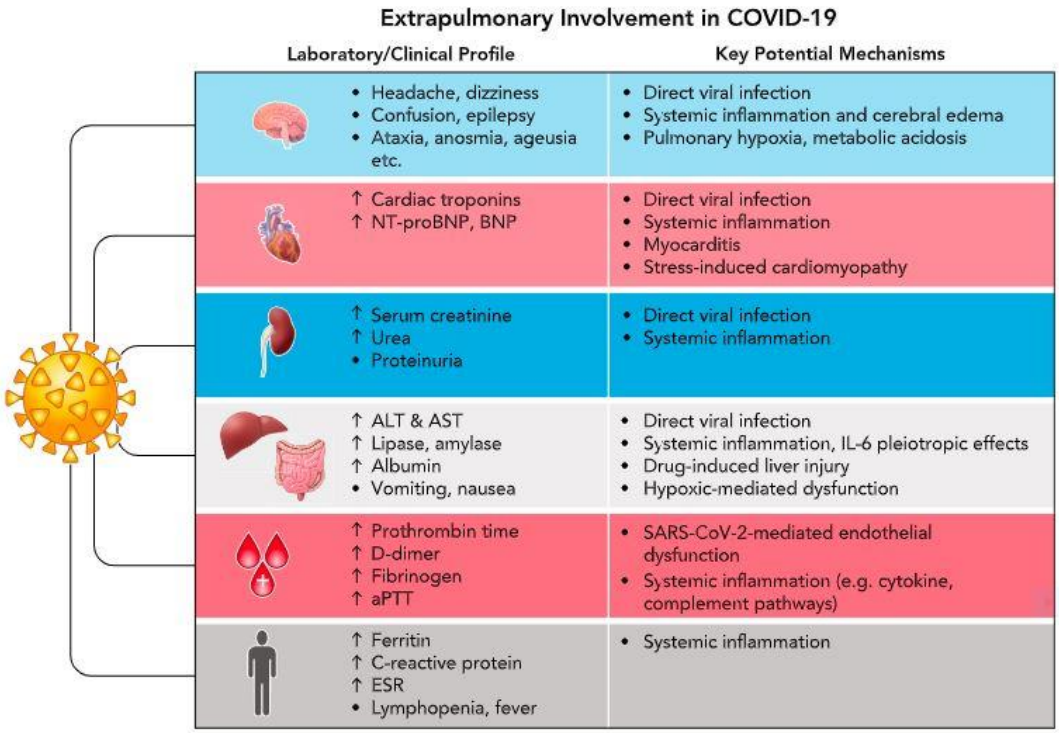


Figure 4. Laboratory/clinical profile and key potential mechanisms underlying extrapulmonary manifestations observed in COVID-19 patients. NT-proBNP, NH2-terminal-proB-type natriuretic peptide; ALT, alanine aminotransferase; AST, aspartate aminotransferase; aPTT, activated partial thromboplastin time; SARS-CoV-2, Severe Acute Respiratory Syndrome Coronavirus 2; ESR, erythrocyte sedimentation rate. From Bohn M.K. et al., 2020.

A timely, localized, and well-coordinated immune response presents the first line of physiological defense against SARS-CoV-2 infection (**Figure 5**). Similar to other cytopathic viruses, SARS-CoV-2 infection induces cellular death and injury in airway epithelial cells through diverse processes such as pyroptosis (Bohn M.K. et al., 2020). Pyroptosis is a highly inflammatory form of programmed cell death that is commonly seen with cytopathic viruses. This is a likely trigger for the subsequent inflammatory response. Viral-mediated cell death causes release of various damage-associated molecular patterns (DAMPs) and pathogen-associated molecular patterns (PAMPs), which are believed to be recognized by pattern-recognition receptors on alveolar macrophages and endothelial cells. For example, Toll-like receptors (TLRs) recognize PAMPs in mostly the extracellular space, triggering induction of pro-inflammatory cytokine transcription factors such as NF-kB, as well as activating interferon regulatory factors that mediate the type I interferon-dependent antiviral response.

In contrast, nucleotide-binding domain leucine-rich repeat (NLR) proteins recognize DAMPs expressed intracellularly, thus triggering activation of inflammasomes and con-version of proIL-1β to active IL-1β. Circulating levels of IL-1β in COVID-19 patients suggests local inflammasome activation with no systemic manifestations. In total, these processes foster an increased secretion of pro-inflammatory cytokines and chemokines, such

as IL-6, type II interferon (IFNγ), monocyte chemoattractant protein 1 (MCP1), and interferon gamma-induced protein 10 (IP-10), as well as subsequent pulmonary recruitment of immune cells, including macrophages and dendritic cells. These cytokines are indicators of a T helper 1 (TH1) cell-polarized response. Secretion of such cytokines and chemokines attracts immune cells, notably monocytes and T lymphocytes, but not neutrophils, from the blood into the infected site. Pulmonary recruitment of immune cells from the blood and the infiltration of lymphocytes into the airways may explain the lymphopenia and increased neutrophil–lymphocyte ratio seen in around 80% of patients with SARS-CoV-2 infection (Tay M.Z. et al., 2020, Bohn M.K. et al., 2020).

Direct viral infection of macrophages and/or dendritic cells is estimated to propagate further cytokine and chemokine release, subsequently activating late-phase immune-cell recruitment of antigen-specific T cells to destroy virally infected alveolar cells. In addition to cytokine release and immune cell recruitment, another potential mechanism that could contribute to successful viral clearance is antibody neutralization. Seroconversion in COVID-19 patients occurs ~7–14 days post symptom onset.

In most individuals, the combined immune response of initial cytokine release and activation of antiviral interferon response followed by immune-cell recruitment should result in successful SARS-CoV-2 clearance from the lungs. However, in some patients, a dysfunctional immune response occurs, which triggers a cytokine storm that mediates widespread lung inflammation.

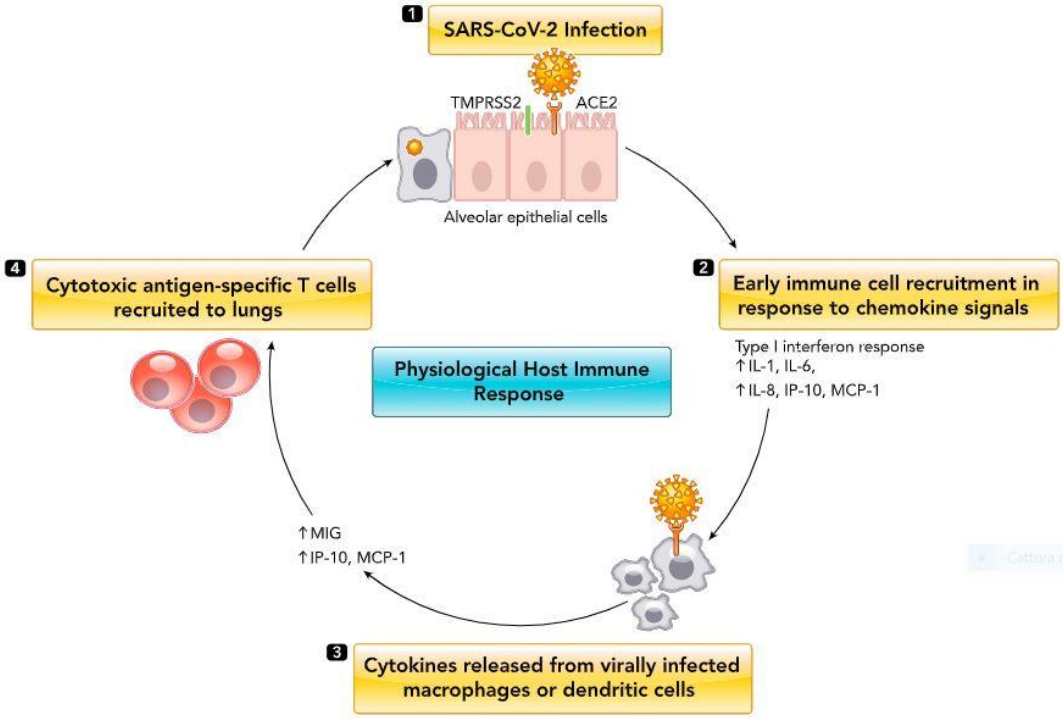


Figure 5. Physiological host immune response to SARS-CoV-2 infection. 1: SARS-CoV-2 enters alveolar epithelial cells by binding to ACE2 through surface S protein mediated by TMPRSS2. 2: pulmonary recruitment of macrophages and dendritic cells in response to chemokine and cytokine release (early phase). 3: direct viral infection of pulmonary macrophages and dendritic cells causes expression of several pro-inflammatory cytokines and chemokines. 4: dendritic cells phagocytose virus in the lungs, migrate to secondary lymphoid organs, and activate antigen-specific T cells, which travel to the lungs and destroy virally infected alveolar cells. From Bohn M.K. et al., 2020.

It was observed that patients with severe COVID-19, requiring intensive care in hospitals, exhibited higher blood plasma levels of IL-6, IL-2, IL-7, IL-10, granulocyte colony-stimulating factor (G-CSF), IP-10, MCP1, macrophage inflammatory protein 1α (MIP1α) and tumour necrosis factor (TNF), suggesting a combined TH1 and TH2 cell response (**Figure 6**). IL-6 levels in these patients continue to increase over time and are relatively more elevated in non-survivors than survivors. In addition to the observed maladaptive cytokine release, elevations in more traditional biochemical markers of acute infection, including C-reactive protein (CRP) and ferritin, as well as continual decreases in lymphocytes and significant elevations in neutrophils, are evident. Patients with severe disease show a significant high percentage of CD14+CD16+ inflammatory monocytes in peripheral blood, that secrete inflammatory cytokines, contributing to the cytokine storm. In addition to local damage, cytokine storm also has ripple effects across the body. Elevated levels of cytokines such as TNF can cause septic shock and multi-organ failure (MOF). These may result in myocardial damage and circulatory failure observed in some patients.

The cytokine storm and/or altered vascular function may be in part at the basis of the coagulopathy that can affect patients with severe disease. Arterial and venous thromboses

are potential consequences of a COVID-19 infection. Increased levels of D-dimers, alongside thrombocytopenia, increased soluble P selectin (reflecting platelet activation), with prolonged prothrombin and partial thromboplastin times, may be interpreted as being linked to this increased risk. Despite anticoagulation, a high number of patients with ARDS secondary to COVID-19 developed life-threatening thrombotic complications (Helms J. et al., 2020; Loo J. et al., 2021).

Older people (over 60 years) and people with co-morbidities are more likely to develop such a dysfunctional immune response that causes pathology and also fails to successfully eradicate the pathogen. In contrast, children tend not to develop severe disease despite being capable of experiencing high viral titres; more than 50% of children experienced mild symptoms or were asymptomatic, with less than 6% of children developing severe symptoms.

The different interaction between SARS-CoV-2 and the host immune system is one of the significant causes of variability in the clinical manifestation of the disease (Oran D.P. et al., 2020; Sharma A. et al., 2020; Zhou F. et al., 2020), (**Figure 6**).

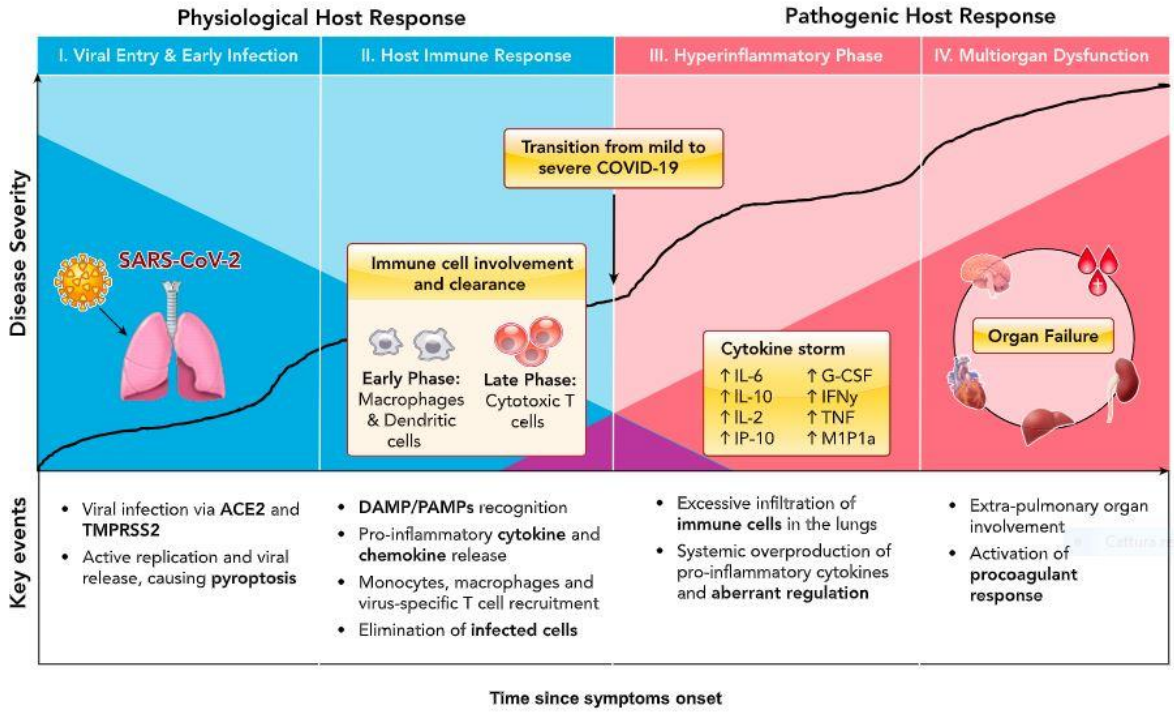


Figure 6. Characterization of key events in COVID-19 disease pathophysiological progression. The dark blue shading indicates physiological viral host response over time, and the dark red shading indicates pathogenic hyper-inflammatory host response over time. From Bohn M.K. et al., 2020.

In some patients, symptoms persist long after the acute phase of the illness. As in acute disease, the most common symptoms of long-COVID are fatigue and dyspnoea,

arthralgia, and chest pain, with around two-thirds reporting more than one symptom. Symptoms may be present in various physiological areas —pulmonary, haematological, renal, endocrine, cardiovascular, neuropsychiatric, gastrointestinal, hepatobiliary, and dermatologic—pointing to the need for a multi-disciplinary approach.

COVID-19 shows a difference in fatality rate between males (2.8%) and females (1.7%)33. As ACE2 is located on the X chromosome, there may be alleles that confer resistance to COVID-19, explaining the lower fatality rate in females. Alternatively, the oestrogen and testosterone sex hormones have different immunoregulatory functions, which could influence immune protection or disease severity (Tay M.Z. et al., 2020).

Finally, COVID-19 manifests itself as a complex disease, with possible multi-organ involvement, and it is therefore necessary to evaluate the effects of the virus on different tissues and body districts. The presence of the ACE2 receptor on different organs is certainly the first cause of the multisystem clinical course of the disease.

2.3 Host genetics as a risk factor.

The COVID-19 pandemic has having a massive impact on public health, societies, and economies worldwide. Despite the vaccination program and treatment have been the high priority, thus, a better understanding of the disease was urgently needed. For almost 3 years now, both scientists and clinicians have been trying to understand why the majority of individuals are asymptomatic, while others undergo life-threatening viral pneumonia and acute respiratory distress syndrome. Age, sex, and comorbidities are relevant clinical variables in determining the response to SARS-CoV-2 infection (Wu C. et al., 2020). However, these risk factors do not explain the severity of COVID-19, particularly in healthy young subjects.

It has been demonstrated that COVID-19 is a complex multifactorial disease, but its main environmental factor (SARS-CoV-2) is easily detectable by a PCR-base swab test. Thus, it represents an accessible disorder for identifying the role of human genetics in susceptibility to infection. Indeed, based on classical twin studies it has already stressed the fact that there is a genetic component associated with the highly varied clinical outcomes of COVID-19. A study, based on data from over 3000 TwinsUK volunteers, found a substantial genetic influence for predicted COVID-19 (heritability of 31%) (Williams F.M.K. et al., 2020). Additionally, de Castro M.V. et al. compared the concordance rate in 10 pairs of young twins, 5 monozygotic (MZ), and 5 dizygotic (DZ), and reported a higher concordance

rate in the MZ group (83%), confirming the significant role of the genetic make-up in the variable clinical manifestations of COVID-19 (**Figure 7**).

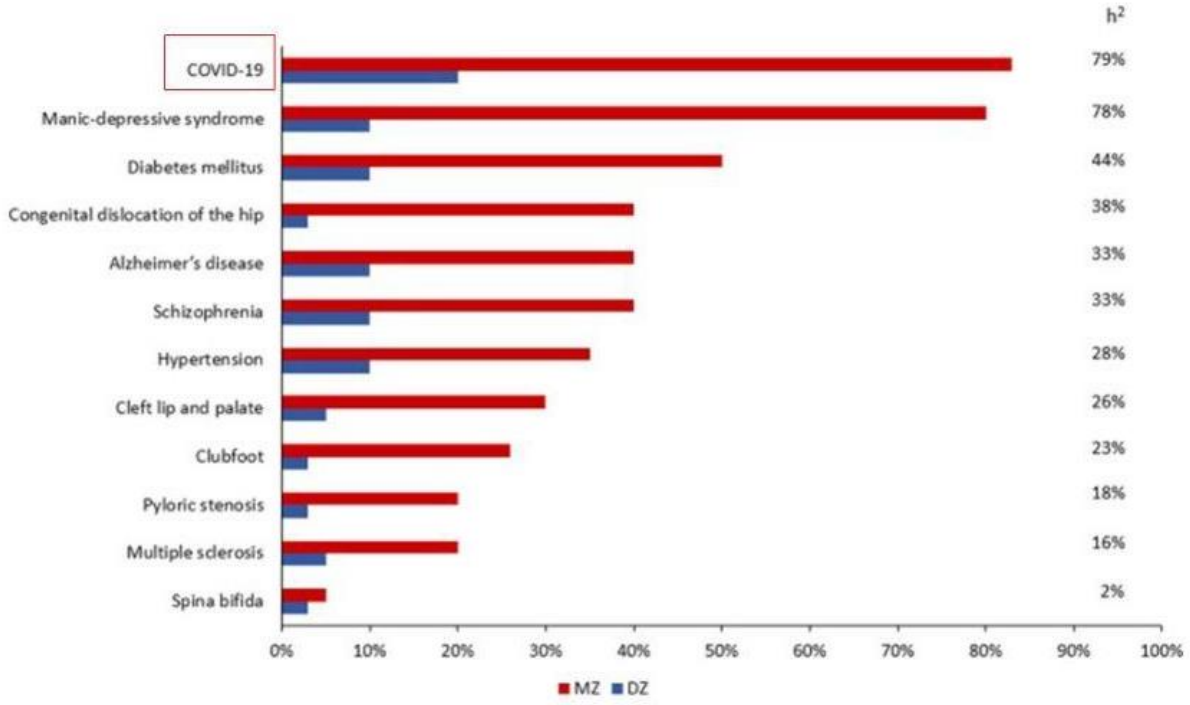


Figure 7. **Twin concordance.** Estimated monozygotic (MZ) and dizygotic (DZ) twin concordance rates for various medical disorders. The percentage referred to the inheritance (h^2) for each condition has been calculated using the formula: h2 = (CMZ – CDZ) / (1 – CDZ). In each multifactorial trait, the concordance rate in MZ twins exceeds that in DZ twins. COVID-19 seems to have a high heritability with a concordance rate of 80% in MZ twins. From Zguro K. et al., 2022.

Based on these results, different methods have been engaged to reveal the genomic factors of COVID-19 susceptibility and severity.

The first traditional approach, trough genome-wide association studies (GWAS), has revealed some common polymorphisms in relevant genes (Severe Covid-19 GWAS Group et al., 2020; Pairo-Castineira E. et al., 2021; Andolfo I. et al., 2021; Shelton J.F. et al., 2021; COVID-19 Host Genetics Initiative, 2021; Kousathanas A. et al., 2022; COVID-19 Host Genetics Initiative, 2022). While, applying the Burden test and focusing only on rare coding variants, it has not been identified any significant associations (Kousathanas A. et al., 2022; Kosmicki J.A. et al., 2021).

On the other hand, it is worth emphasizing the fast-growing role of machine learning (ML) models in classification or clustering tasks in genomic datasets (Molla M. et al., 2004). This last model will help in resolving the genetic variation underlying COVID-19 by combining rare and common variants into an overall predictive model.

GWAS study single-nucleotide polymorphisms (SNPs) that are reported as clusters of correlated variants demonstrating a statistically significant association with complex disorders. Usually variants fall in non-coding regions of the genome and therefore do not directly affect the coding sequence of a gene. These studies require sample sizes of ten/hundred thousand subjects to have sufficient statistical power to detect a moderate association while analysing hundreds of thousands to millions of SNPs Predominately, GWAS focus mainly on common variants, usually with a minor allele (MAF) ≥5% whose effects are relatively small. A major limitation of genome-wide approaches is the necessity to choose a high level of significance, $p < 5 \times 10^{-8}$, because of the multiple independent tests.

International collaborations allowed doing GWAS for COVID-19, thanks to sharing scientific methods and resources, to shed light on the genetic factors of SARS-CoV-2 infection and the outcomes of the resulting disease. Up to date, multiple studies have successfully identified various significant loci associated with susceptibility and/or severity of COVID-19 (Severe Covid-19 GWAS Group et al., 2020; Pairo-Castineira E. et al., 2021; Andolfo I. et al., 2021; Shelton J.F. et al., 2021; COVID-19 Host Genetics Initiative, 2021; Kousathanas A. et al., 2022; COVID-19 Host Genetics Initiative, 2022). However, these variants only explain a small fraction of trait variability and, as it is well documented, GWASs are difficult to interpret because they often associate non-coding variants with phenotype; therefore, the relevant genes need to be pinpointed by deeper follow-up analyses.

Because of GWASs focus on identifying common variants, it is possible that the analysis of rare variants (MAF < 0.5%) could further add more elements to clarifying the role of rare genetic factors in the mechanism of SARS-CoV-2 infection. For this purpose, the Burden test is another robust and traditional method (Guo M.H. et al., 2018). This method consists in an aggregation of rare, protein-altering variants and a comparison between case and control subjects. Similar to GWAS, in order to detect statistically significant associations, the Burden test needs hundreds of thousands of subjects. Until now, different working groups have tried to characterize rare variants and clarify the biological mechanism of severe COVID-19, but without significant evidence of association yet (Kousathanas A. et al. 2022; Kosmicki J.A. et al., 2021;).

2.4 COVID-19 as a complex disease

Because of some tens of genes, discovered both by GWAS and Burden tests, were not sufficient to explain the heritability of the disease and to fully predict severity, new approaches able to identify the entire genetic variability, and combine both common and rare variants are necessary.

Concerning complex diseases, the height is the prototype of polygenicity (Boyle E.A. et al., 2017). Such complex traits are products of many genes which interact together in a complex way. Hundreds of common variants, as well as rare and low-frequency variants, have been reported to be associated with height (Boyle E.A. et al., 2017; Marouli E. et al., 2017). The disease risk is mostly determined by genes not directly relevant to the disease and by a much smaller number with direct effects (Boyle E.A. et al., 2017). Fort this reason genetic features for complex traits could reach thousands or even hundreds of thousands.

Therefore, COVID-19 is a complex multisystem disorder, and as such, a much greater number of genes are expected to be involved; much greater than those reported by GWAS and Burden tests.

Methods neglecting the combined contribution of common and rare variants were unlikely thus far to thoroughly characterize the host genetics underlying COVID-19. Thus, new ML methods are under development.

The methods applied so far did not consider the combined contribution of common and rare variants in order to comprehensively characterize the host genetics underlying COVID-19. The application of machine learning (ML) models allows to combine all the genetic variability of an individual in a single model, identifying variants associated with disease severity (Zhang S. et al., 2021).

The main concept of ML algorithms is to automatically learn relevant information from data. Performance generally improves as the number of samples increasing. The availability of huge genomic datasets has led to the increasing application of ML techniques in clustering tasks related to many biological and medical fields. Supervised learning (SL) and unsupervised learning (UL) are examples of two different types of ML model approach. They differ in the way the models are trained and the condition of the training data that's required. Each approach has different strengths, so the task or problem faced by a supervised vs unsupervised learning model will usually be different (Sidey-Gibbons J.A.M. et al., 2019).

3. Aim of the study

The aim of this study is to describe and highlight the genes that we have identified to be associated with susceptibility or severity of COVID-19, using a new predictive model such as the post-Mendelian model. We proposed this new model for a genetic characterization of the disorder based on an adapted Poly-genic Risk Score (PRS), called Integrated PolyGenic Score (IPGS). This allowed us to reach a more precise disease severity prediction than that based on sex and age alone.

Outlining virus-host interactions is critical to elucidate the pathogenesis of COVID-19 and to translate these findings in patient care improvement and further drug development. Furthermore, advances in modelling the interplay between SARS-CoV-2 and host genetics hold significant promise for addressing other complex diseases.

4. Materials and methods

4.1 Study design and patient enrolment

The enrolment of patients included in this study began in March 2020 and is still ongoing.

The GEN-COVID Multicenter Study was founded by Prof. Alessandra Renieri and it started its activity on March 16, 2020, following approval by the Ethical Review Board of the Promoter Center, University of Siena (Protocol n. 16917, approval dated March 16, 2020). To date, this Multicenter Study includes a network of more than 30 Italian hospitals, 16 local healthcare units and 8 departments of preventative medicine (https://drive.google.com/open?id=13bkMkYBLApIMrSwoj9fJ30v3dCiFwsSh&usp=sharing).

The aim of the Multicenter group was to collect and systematize biological samples and clinical data across multiple hospitals and healthcare facilities in Italy with the purpose of deriving patient-level phenotypic and genotypic data, and the specific intention to make samples and data available to COVID-19 researchers globally. To reach these aims, the project collected and organized high-quality samples and data whose integrity was assured and could be readily accessed and processed for COVID-19 research using existing interoperability standards and tools. For this reason, a GEN-COVID Biobank (GCB) and a

GEN-COVID Patient Registry (GCPR) were established utilizing already existing biobanking and patient registry infrastructure.

The purpose of the GEN-COVID Multicenter Study was to recruit at least 2000 positive subjects, with a wide range of disease severity, rang from hospitalized patients with severe COVID-19 disease to asymptomatic individuals, and collect clinical and genetic data in order to identify potential links between genetic variants and clinical variability, patient presentation, and disease severity. To achieve this overall aim, it has been performed Whole Exome Sequencing (WES) by the University of Siena. At the same time, thanks to the collaboration with other International Study Group such as *The COVID-19 Host Genetics Initiative* (HGI), it has been possible to insert our cohort in GWAS studies; the analysis for our cohort was performed at the Institute for Molecular Medicine of Finland (FIMM). The WES data have been shared through the Network of Italian Genome (NIG, http://www.nig.cineca.it/ NIG database, http://nigdb.cineca.it) at CINECA, the largest Italian computing center.

In order to ensure a collection that could be, as much as possible, comprehensive and representative of the Italian population, hospitals from across Italy, local healthcare units, and departments of preventive medicine have been involved in collecting samples and associated patient-level data.

At first, the inclusion criteria for the study are PCR-positive SARS-CoV-2 infection, age ≥18 years, and appropriately given informed consent that includes detailed information about the study, maintaining the confidentiality of personal data. Then we decided also to enrol resistors and paediatric subjects. In addition to the samples collection, an extensive questionnaire is used to assess disease severity and collect basic demographic information from each subjects.

In particular, the questionnaire included socio-demographic information such as sex, age, and ethnicity, information about family history, (pre-existing) chronic conditions, and more than 150 items concerning to SARS-CoV-2 related symptoms synthesized in a binary mode for each involved organ/system: lung, heart, liver, pancreas, kidney, and olfactory/gustatory and lymphoid systems. It is important to specify that these clinical data were continually updated as new information appeared regarding COVID-19.

Peripheral blood samples in ethylenediamine tetraacetic acid-containing tubes were collected for all subjects. Genomic DNA was centrally isolated from peripheral blood samples using the MagCore® Genomic DNA Whole Blood Kit (Diatech Pharmacogenetics, Jesi, Italy) according to the manufacturer's protocol at the Promoter Center. For all subjects,

aliquots of plasma and serum are also available. Whenever possible, leukocytes were isolated from whole blood by density gradient centrifugation and stored in dimethyl sulfoxide solution and frozen using liquid nitrogen. For the majority of cohort, swab specimens are also available and stored at the reference hospitals

Peripheral blood samples in ethylenediamine tetraacetic acid-containing tubes were collected for all subjects. Genomic DNA was centrally isolated from peripheral blood samples using the MagCore® Genomic DNA Whole Blood Kit (Diatech Pharmacogenetics, Jesi, Italy) according to the manufacturer's protocol. For all subjects, aliquots of plasma and serum are also available. Whenever possible, leukocytes were isolated from whole blood by density gradient centrifugation and stored in dimethyl sulfoxide solution and frozen using liquid nitrogen. For the majority of cohort, swab specimens are also available and stored at the reference hospitals.

The clinical severity of COVID-19 was assessed using a modified version of the WHO COVID-19 Outcome Scale (WHO R&D Blueprint Novel Coronavirus COVID-19 Therapeutic Trial Synopsis, 2020), identifying the following seven categories:

-1 = "resistors" (cases in which a single person in a family remains uninfected while everyone around them gets sick; or people who work in a high-risk environment but who do not get sick);

0 = not hospitalized, asymptomatic or oligo-symptomatic;

1 = hospitalized, not receiving supplemental oxygen;

2 = hospitalized, receiving low-flow supplemental oxygen;

3 = hospitalized, receiving continuous positive airway pressure (CPAP) or bi-level positive airway pressure (BiPAP) ventilation;

4 = hospitalized, receiving invasive mechanical ventilation;

5 = deceased.

4.2 WES and GWAS

Advances in next-generation sequencing have led to the production of a huge amount of genome sequence data that allows for a more complete understanding of human genetics. WES is an efficient method that offers high coverage with a low cost. This Sequencing targets approximately 3% of the whole genome, which is the basis for protein-coding genes. (Suwinski P. et al., 2019).

WES with at least 97% coverage at 20x was performed using the Illumina NovaSeq6000 System (Illumina, San Diego, CA, USA). Library preparation was performed using the Illumina Exome Panel (Illumina) according to the manufacturer's protocol. Library enrichment was tested by qPCR, and the size distribution and concentration were determined using Agilent Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA, USA). The Novaseq6000 System (Illumina) was used for DNA sequencing through 150 bp paired-end reads. Variant calling was performed according to the GATK4 (O'Connor and Auwera 2020) best practice guidelines, using BWA (Li and Durbin 2010) for mapping and ANNOVAR (Wang K et al., 2010) for annotating.

Genotyping data on 700,000 genetic markers were obtained on genomic DNA using the Illumina Global Screening Array (Illumina) according to the manufacturer's protocol. Homo sapiens (human) Genome Reference Consortium Human Build 38 (GRCh38) was used. Quality checks (SNP calling quality, cluster separation, and Mendelian and replication error) were done using GenomeStudio analysis software (Illumina). The computer package Plink v1.90 was used to process 700k SNP-genotyping data and to calculate SNP genotype statistics.

4.3 Definition of the Boolean features

For each patient, the average number of variants that are called in a WES analysis stands at around 20,000. In order to simplify and decrease the number of variants to be used for the overall analysis of the ML data, all the variants obtained from the sequencing were converted into 12 sets of Boolean features to better represent the variability at the gene level, adopting the *gene-based* method, already used in other association studies (Hägg S. et al. 2015; Chung J et al. 2019).

First, any variant not impacting on the protein sequence was discarded. Then, the remaining variants were classified according to their minor allele frequency (MAF) as reported in gnomAD for the reference population as: ultra-rare, $MAF < 0.1\%$; rare, $0.1\% \leq MAF < 1\%$; low-frequency, $1\% \leq MAF < 5\%$; and common, $MAF \geq 5\%$.

Non-Finnish European (NFE) was used as a reference population. SNPs with MAF not available in gnomAD were treated as ultra-rare. INDELs with frequency not available in gnomAD were treated as ultra-rare when present only once in the cohort and otherwise discarded as possible artefacts of sequencing.

For the ultra-rare variants, 3 alternative Boolean representations were defined, which are designed to capture the autosomal dominant (AD), autosomal recessive (AR), and X-linked (XL) model of inheritance, respectively.

The AD and AR representations included a feature for all the genes on autosomes. These features were equal to 1 when the corresponding gene presented at least 1 for the AD model, or 2 for the AR model, variants in the ultra-rare frequency range and 0 otherwise.

The XL representation included only genes belonging to the X chromosome. These features were equal to 1 when the corresponding gene presented at least 1 variant in the ultra-rare frequency range and 0 otherwise.

The same approach was used to define AD, AR, and XL Boolean features for the rare and low-frequency variants. Common variants were represented using a different approach that is designed to better capture the presence of alternative haplotypes. For each gene, all the possible combinations of common variants were computed.

For instance, in the case of a gene belonging to an autosome with 2 common variants (named A and B), 3 combinations are possible (A, B, and AB), and (consequently) 3 Boolean features were defined both for the AD and AR model. In the AR model each of these 3 features was equal to 1 if all the variants in that particular combination were present in the homozygous state and 0 otherwise. The same rule was used for the AD model, but setting the feature to 1 even if the variants in that particular combination are in the heterozygous state. In both models, AD and AR, a further feature was defined for each gene to represent the absence of any of the previously defined combinations. In the AD model, this feature was equal to 1 if no common variant is present and 0 otherwise; in the AR model, it is equal to 1 if no common variant is present in the homozygous state and 0 otherwise. The same approach was used to define the set of Boolean features for common variants in genes belonging to the X chromosome (Fallerini C. et al., 2022).

4.4 PC analysis

The genetic ancestry of the patients was estimated using a random forest classifier. SNPs of autosomes with MAF above 10% and in linkage disequilibrium were extracted from the 1000 genome project using BCFTOOLS (Danecek et al., 2021), and intersected with variants from our cohort of COVID-19 patients. The resulting set of variants was used to compute 6 Principal Components by PLINK (Purcell et al. 2007) using samples from the *1000 genome projects*. GEN-COVID samples were projected along the same Principal

Components. The Random Forest classifier, as implemented in the *R library randomForest* (Liaw and Wiener 2002), was trained using samples from 1000 genomes with known ancestry, and then used to predict ancestry for the cohort. To avoid *bias* in the analysis due to the different ethnicity, only patients of genetic European ancestry were retained for further analyses

4.5 Machine learning and statistical analysis

In order to study the role of host genetics in the severity of COVID-19, we applied a ML approach that would allow us to identify the most relevant genes/variants involved and at the same time be able to predict the severity of the disease based on the genetic information extracted from the WES. The subsets of the most relevant features were identified using logistic regression models with Least Absolute Shrinkage and Selection Operator (LASSO) regularization

The variants classification was done comparing cases, i.e. patients with severe disease, versus controls, i.e. patients with mild disease. The coefficients of the LR model are directly related to the degree of importance of the corresponding variants. The aim of LASSO regularization is to minimize (shrink) the number of coefficients of the model, consequently minimizing the number of input features used for predicting outcomes, reducing *overfitting.* In this classification, positive coefficients indicate variants associated with COVID-19 severity, whereas negative coefficients indicate protective variants.

As already reported, only variants that have a protein impact were selected for Boolean representation. This allowed to reduce the number of variants and to improve the interpretability regarding the biological significance of the extracted variants. The chi-square test or the Mann Whitney test (for continuous variables) were used to assess the statistical association between the clinical severity of the disease and clinical variables such as sex, ethnicity, comorbidities and possible involvement of other organs or systems.

Ad hoc functional experiments were performed to demonstrate the biological impact of a specific variant on the protein and interpret the biological significance towards COVID-19 disease.

5. Results

5.1 Patients' clinical classification corrected for sex and age

The role of gender and age in the COVID-19 severity has been widely demonstrated, defining male and older individuals (> 80 years) as most at risk.

In order to obtain a clinical classification as independent as possible from age and gender, an Ordered Logistic Regression (OLR) model was used. Separately for the male and female cohorts, two OLR models were fitted using the age to predict the ordinal grading (0, 1, 2, 3, 4, 5) dependent variable. Then, each patient had clinical classification equal to: 0 (mild), if the actual patient grading was below the one predicted by the OLR; or 1 (severe), if the grading was above the OLR prediction. The patients with a predicted gradient equal to the actual gradient were excluded from the LASSO analysis, by which we wanted to compare the "extreme ends" (**Figure 8**), (Picchiotti N. et al., 2021). This procedure allowed us to divide our cohort into cases (OLR = 1) and controls (OLR = 0). This classification allowed us to use SARS-CoV-2 positive subjects with a mild phenotype as controls. This procedure also has the advantage of isolating patients whose genetic factors are more important for predicting the severity of COVID-19, compared to gender and age.

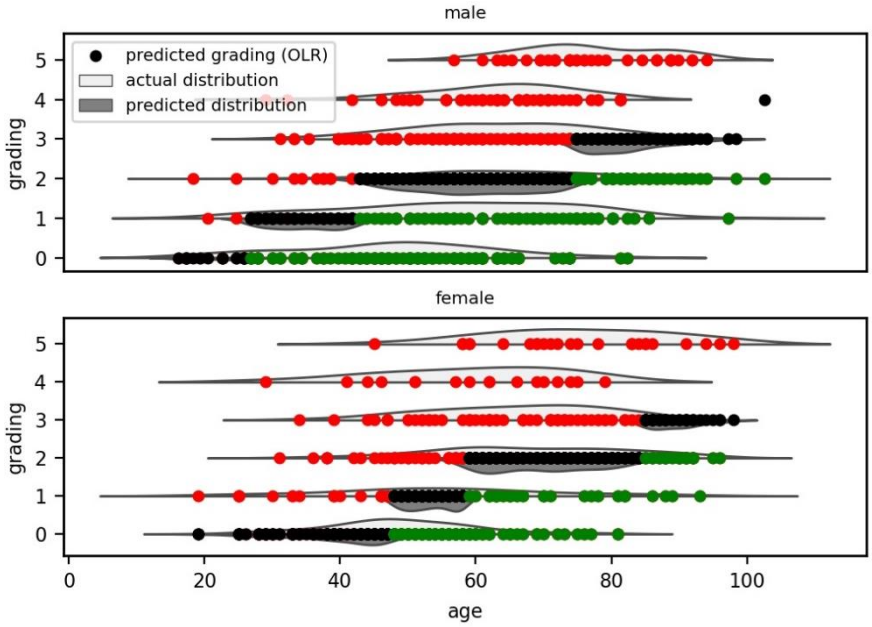


Figure 8. **Clinical classification adjusted by age.** Ordered Logistic Regression (OLR) model for male and female. On the Y axis, the grading according to patients' treatment is reported (5=deceased; 4=intubated; 3=CPAP/biPAP; 2=oxygen therapy; 1=hospitalized without oxygen support; 0=not hospitalized oligo-asymptomatic patients). On the X axis, age is reported. Red dots represent subjects falling above the expected treatment according to age (hence considered severe); green dots are subjects falling below the expected treatment according to age (hence considered mild); black dots are subjects matching the expected treatment according to age (hence considered intermediate). From Picchiotti N. et al., 2021.

5.2 Features' selection

The extraction of features associated with the genetic basis of COVID-19 severity was obtained by LASSO LR model, as already described in methods. The patients with a predicted grading equal to the actual grading were excluded. The remaining patients were divided into two classes depending on whether their actual phenotype was milder (OLR = 0) or more severe (OLR = 1) than the one expected for a patient of that age and sex. Patients classified as 1 were considered as cases and patients classified as 0 were considered as controls for the 12 LASSO logistic models based on the 12 separate Boolean representations. From the initial 163,099 cumulative features in 12 Boolean representations, the selected features contributing to COVID-19 clinical variability were 7249 (4,4%). Around 25% of the genes were sex-specific. (**Figure 9**), (Fallerini C. et al., 2022).

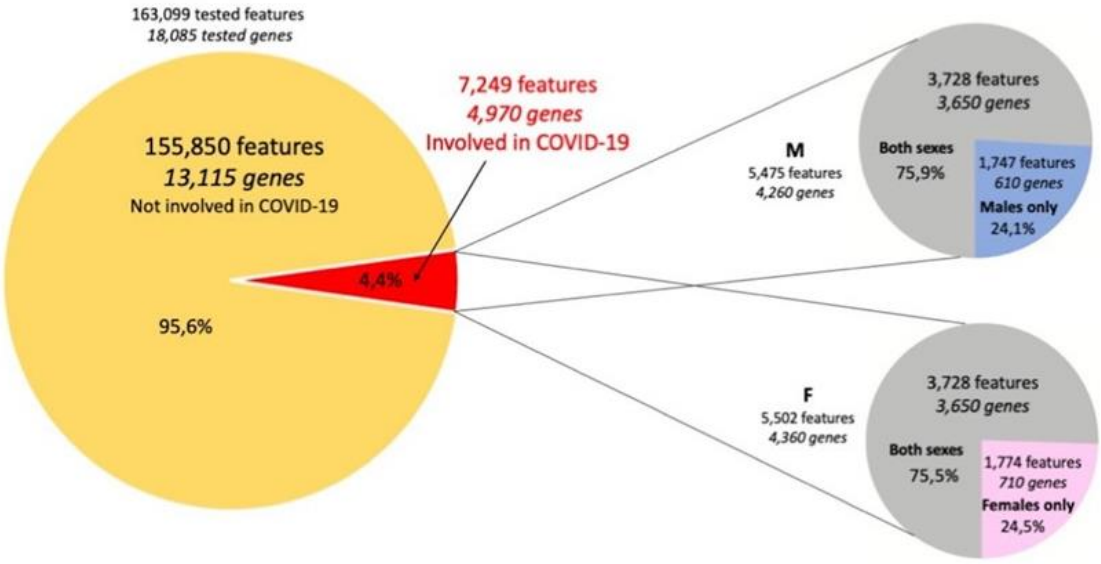


Figure 9. **Features' selection**. From the initial 163,099 cumulative features (divided into 36,540 ultra-rare, 23,470 rare, 13,056 low frequency and 90,033 common features) in 12 Boolean representations, the selected features contributing to COVID-19 clinical variability are 7249. The total number of genes contributing to COVID-19 clinical variability was 4260 in males and 4360 in females, 75% of which were in common and 25% were sex-specific. Modified from Fallerini C. et al., 2022.

5.3 Biological interpretability of extracted genetic features

Selected genes contributed by ultra-rare, rare, low-frequency variants, or/and common variants. Specifically, 54% contributed by only one, 29% by two, 11% by three, and 6% by four types of variants. Around 25% of the genes were sex-specific. This last group

consist of either gene located on the X chromosome, such as *TLR7* and *TLR8* in males, or genes regulated by sex hormones, such as *AR*, *TLR3*, *SELP* and *ADAMTS13*. It was found a 26% matching between the extracted genes and the already known genes associated with susceptibility to viral infections. We further investigated extracted genes using the *Human Gene Connectome* (HGC) to measure their connection with known genes related to viral susceptibility and disease severity. Interestingly, of the 4943 genes of our model that are mapped in HGC, 4401 (89%) are biologically significantly related ($p < 0.05$) and 2847 (57.6%) to a degree of 0 (overlap) or 1 (direct connection) with one of the genes of the three Core Lists, *Virus_Interactome*, *PanelApp* e *GWAS_Functional.* Among the extracted ultra-rare variants there was a group of genes, such as *TLR3, TLR7* and *TICAM1*, already shown to be directly involved in the Mendelian-like forms of COVID-19. Moreover, another group of genes are natural candidates because of their function: these include the *ACE2, ADAM17*, *TLR5*, *SLC26A9*, CFTR-related genes, genes involved in glycolipid metabolism, genes expressed by cells of the innate immune system, and genes involved in the coagulation pathway. Finally, a group of genes such as *ACE2* (if involved by ultra-rare variants) confers protection from the severe disease. This group includes several genes whose mutations are responsible for auto-inflammatory disorders.

Among the low-frequency variants extracted, we identified some genes associated with either severity or protection from severe COVID-19 that are linked to the CFTR pathway (*PSMA6*) as well as specific genes involved in the immune response (*NOD2*).

The model was also able to identify a group of extracted common variants already shown to be linked to either severe or mild COVID-19, such as the L412F *TLR3* and D603N *SELP* polymorphisms, already reported to be associated with the severe disease.

Several coding polymorphisms in Linkage Disequilibrium (LD) with already reported genomic SNP were extracted, such as the ABO blood group, *OAS1-3* genes, *PPP1R15A* gene and others (Elhabyan A. et al., 2020).

In conclusion, considering their functions, genes involved in the immune and inflammatory responses, or those involved in the coagulation pathway and NK and T cell receptor, are to be considered natural candidates for severe or mild COVID-19.

We then investigated what was the impact of some of the extracted genes in determining the disease. In particular, we first considered the genes that contributed with rare or ultra-rare variants, in the idea of defining the existence of a Mendelian form of the disease and subsequently we investigated the genes that contributed with common variants

in the idea of defining the pathogenic mechanism associated with the combination of multiple polymorphisms in the polygenic form of the disease.

5.4 COVID-19: a Mendelian monogenic disease

Although disease severity is disproportionally higher among the elderly, men and individuals with chronic comorbidities, severe or critical cases of COVID-19 can also occur in younger, previously healthy individuals. Since the risk of critical COVID-19 is most significantly correlated with age, its manifestation in relatively young patients could indicate the presence of strong predisposing genetic variants that significantly impair the core immune pathways engaged in the defense against SARS-CoV-2 infection. Additionally, elderly individuals could be identified with similar genetic variants that until now had been redundant in their immune response, but became manifest in the setting of SARS-CoV-2 infection. While these predisposing host genetic variants may be individually rare and explain a minority of severe or critical COVID-19 cases, the identification of such monogenic disorders can help to elucidate the mechanistic basis of the immune-pathogenesis underlying severe COVID-19 and to facilitate preventive, diagnostic and personalized therapeutic interventions (van der Made C.I. et al., 2022).

5.4.1 X-linked COVID-19 due to *TLR7* rare variants

The first novel immunodeficiency discovered in patients with critical COVID-19 was the X-linked TLR7 (Toll-like Receptor 7) deficiency (van der Made C.I. et al., 2020). In two unrelated young brother pairs with critical COVID-19, rare genetic variants in the X-chromosomal *TLR7* were identified by rapid clinical WES. This gene encodes an evolutionary highly conserved cytosolic pattern recognition receptor that recognizes single-stranded RNA viruses such as coronaviruses (Casanova J.L. et al., 2011). Previously, it had been revealed that mice deficient in either *TLR7* or the downstream adaptor MyD88 exhibited impaired production of IFN-I, delayed viral clearance and severe lung pathology upon infection with MERS-CoV (Cervantes-Barragan L. et al., 2007; Channappanavar R. et al., 2019; Zhao J. et a., 2014). *TLR7* is most abundantly expressed on plasmacytoid dendritic cells (pDCs), which are important producers of type I IFN. In peripheral blood cells isolated from these young male patients, the *TLR7* variants were shown to lead to an absence of the

transcriptional IFN-I response and the production of interferon gamma (IFNγ) in response to a TLR7-specifc agonist (van der Made C.I. et al., 2020).

Applying the LASSO logistic regression analysis, after correcting for Principal Components, to a synthetic Boolean representation of the entire set of genes of the X chromosome on the extreme phenotypic ends of the male subset of the Italian GEN-COVID cohort, *TLR7* was picked up as one of the most important susceptibility genes for severe disease (Fallerini C. et al., 2022).

Overall, by selecting for young (<60 year-old) males the association between the presence of *TLR7* rare variants and severe COVID-19 was significant (p=0.037 by Fisher Exact test) (Fallerini C. et al., 2021). *TLR7* missense variants impacts on protein function (CADD > 12.28) in 5 out of 79 male patients (6.3%) with life-threatening COVID-19 (hospitalized intubated and hospitalized CPAP/BiPAP) and in none of the 77 SARS-CoV2 infected oligo-asymptomatic male participants.

In the entire male cohort of 561 COVID-19 patients (261 cases and 300 controls) regardless of age, *TLR7* rare missense variants have been found in three additional patients over 60 years of age, including two cases (who shared the p.Ala1032Thr variant) and one control, bearing the p.Val222Asp variant, predicted to have a low impact on protein function (CADD of 5.36).

In order to functionally link the presence of the identified *TLR7* missense variants and the effect on the downstream type I IFN-signalling, a gene expression profile analysis has been performed in peripheral blood mononuclear cells (PBMCs) isolated from patients following recovery, after stimulation with the TLR7 agonist imiquimod.

Fold change in mRNA expression of *TLR7* and type 1 IFN-related genes, such as *ISG15, IRF7, IFN-α* and *IFN-γ,* induced by TLR7 agonist imiquimod was compared. This analysis showed a statistically significant decrease of all TLR7-related genes for two variants (Ser301Pro and Ala1032Thr) identified in three cases compared with healthy controls demonstrating a complete impairment of TLR7 signalling pathways in response to TLR7 stimulation. These variants have to be considered loss of function (LOF) variants (Fallerini C. et al., 2021), (**Figure 10, panel a-b**).

In order to validate the functional effect of *TLR7* variants, transfection experiments in HEK293 cells have been performed, cloning a dedicated TLR7 plasmid for each of them. The expression of IFN- *α* in imiquimod stimulated and unstimulated cells by qRT-PCR has been evaluated, confirming the results obtained in PBMCs for the screened variants (**Figure 10, panel c**).

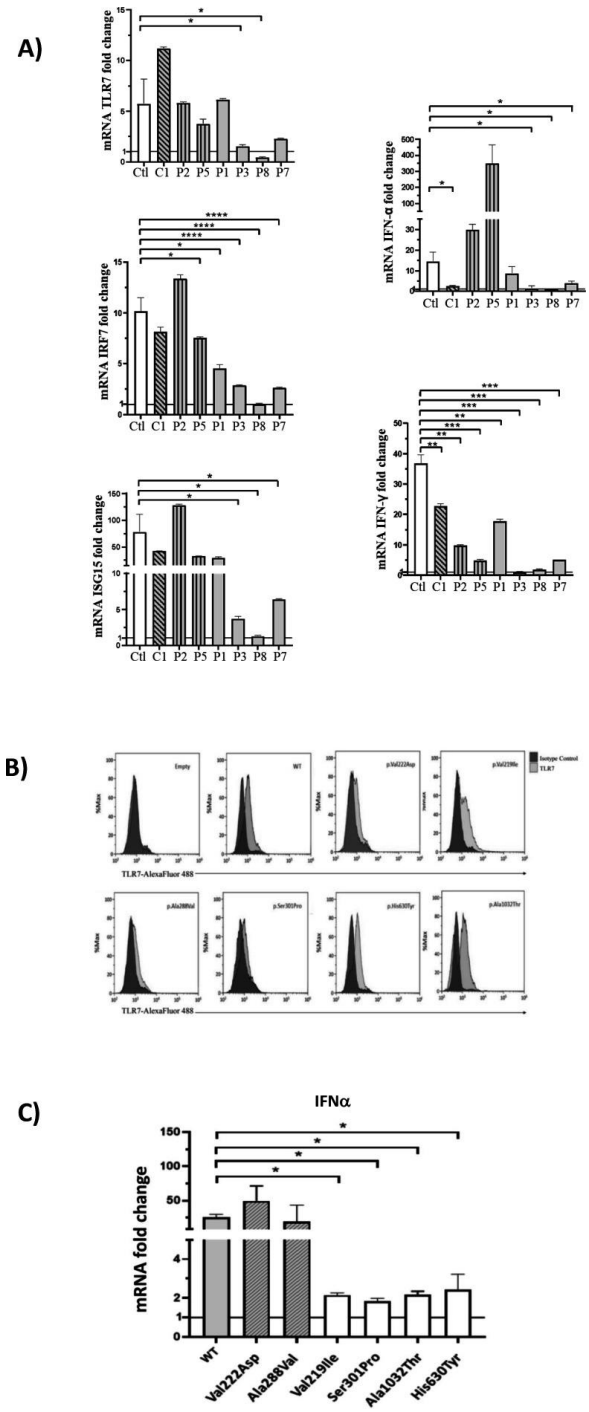
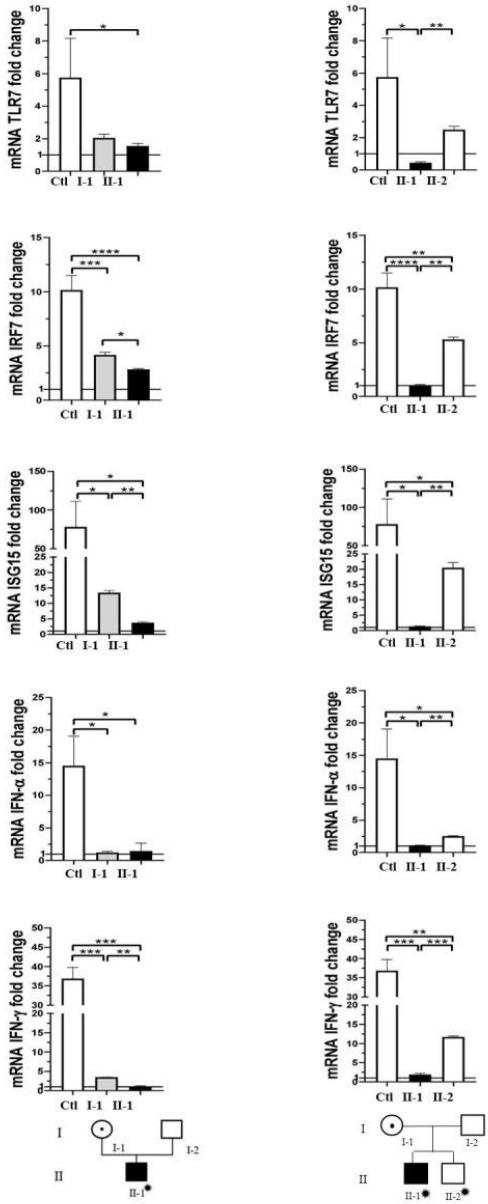


Figure 10. Gene expression profile analysis in PBMCs and in HEK293 cells transfected with the functional variants after stimulation with a *TLR7* agonist. A) PBMCs from COVID-19 patients and six unaffected male and female controls were stimulated with the TLR7 agonist imiquimod or cell culture medium. Ctl indicates healthy controls (white bar); C1, the asymptomatic mutated control (diagonal lines bar); P2, P5, cases with neutral variants (vertical lines bar); P1, P3, P8, P7 cases with functional variants (grey bar). B) Histograms of intracellularly expressed TLR7 protein in HEK293 cells transfected with the different TLR7 plasmids. C) Gene expression profile analysis of IFN-α in transfected cells after stimulation with the TLR7 agonist imiquimod. WT indicates cells transfected with WT *TLR7* plasmid. From Fallerini C. et a., 2021.

For two cases segregation analysis was available: in the two pedigrees, the disease nicely segregated as an X-linked disorder conditioned by environmental factors, that is SARS-CoV-2. This was also supported by functional analysis on all TLR7-related genes. In particular, expression profile analysis in male mutated confirmed a statistically significant reduction compared to the wild-type brother. Moreover, only the infected mutated male had severe COVID-19, whereas the infected not mutated brother was asymptomatic (**Figure 11**).

A



B

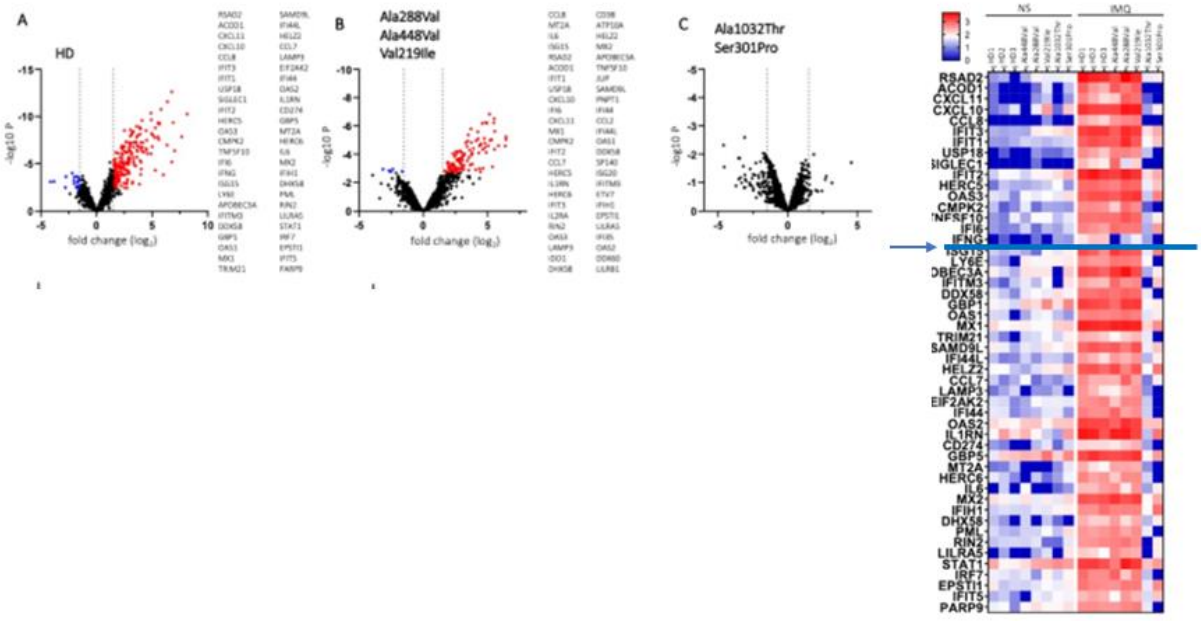
C

ID	Pedigree position	Gender	Age	COVID-19 Clinical Category	Signs/Symptoms	PCR on Naso-pharyngeal Swab	*TLR7* p.Ser301Pro
P3-1	I-1	F	71	/	/	/	Heterozygous
P3	II-1	M	46	3	Bilateral pneumonia, CPAP treatment	+	Hemizygous

ID	Pedigree position	Gender	Age	COVID-19 Clinical Category	Signs/Symptoms	PCR on Naso-pharyngeal Swab	*TLR7* p.Ala1032Thr
P8	II-1	M	66	3	Bilateral pneumonia, CPAP treatment	+	Hemizygous
P8-1	II-2	M	59	0	Asymptomatic	+ at serological	Wild-type

Figure 11. **Segregation analysis.** Fold change in mRNA expression following Imiquimod stimulation of TLR7 itself and its main effectors, IRF7, ISG15, IFN-alpha, and IFN-gamma is shown in Panel A. Grey columns represent individuals harbouring the *TLR7* variant and black columns are severely affected SARS-CoV-2 cases. Pedigree (Panel B) and respective segregation of *TLR7* variant and COVID-19 status (Panel C) are also shown. Squares represent male family members; circles, females. Individuals infected by SARS-CoV-2 are indicated by a virus cartoon close to the individual symbol. From Fallerini C. et al., 2021.

Additionally, RNA sequencing was performed to investigate transcriptome variations following imiquimod stimulation of PBMC isolated from patients carrying previously identified hypo-morphic (p.Ala288Val and p.Ala448Val), hypo-functional (p.Val219Ile), and loss-of-function (p.Ala1032Thr and p.Ser301Pro) *TLR7* variants. It has been revealed a profound impairment of the TLR7 pathway in patients carrying LOF variants. Of note, a failure in IFNγ upregulation following stimulation was also observed in cells harbouring the hypo-functional and hypo-morphic variants. Overall, the transcriptomic profile of cells harbouring LOF *TLR7* variants showed a wide deficiency of IFN-stimulated genes while both hypo-morphic and LOF mutations displayed a reduction of IFNγ transcription (Mantovani S. et al., 2022) (**Figure 12**).



F**igure 12**. **Differentially expressed genes (DEGs) in PBMC from healthy donors (HDs) and patients carrying *TLR7* variants stimulated with IMQ.** A–C Volcano plots showing DEGs. Red dots show upregulated genes and blue dots represent downregulated genes. A) Three HDs. B) Patients (n = 3) carrying the Ala288Val, Ala488Val, and Val219Ile variants. C) Patients carrying Ala1032Thr (n = 1) and Ser301Pro (n = 1) variants. Gene Ontology biological process terms significantly overrepresented among the genes increased by IMQ are shown in the lower panel in (A, B, D). Heatmap of logCPM values for the top DEGs in HDs and patients carrying TLR7 variants after IMQ stimulation. IMQ imiquimod, NS non stimulated. From Mantovani S. et al., 2022.

Overall, these data expand previous findings on the *TLR7* role in rare Mendelian forms of COVID-19 and provide further insights into the altered pathways that might contribute to disease severity. Specifically, missense deleterious variants in the X-linked recessive *TLR7* gene may represent the cause of disease susceptibility to COVID-19 in up to 2% of severely affected young male cases (van der Made C.I. et al., 2020; Fallerini C. et al., 2021; Mantovani S. et al., 2022). Concerning therapeutic approach, it has been shown that at around day 10 in subjects with COVID-19, IFN-I decreased while IFNγ remained stable, promoting the resolution of lung inflammation. Therefore, administration of IFN-I might be considered a therapeutic option for *TLR7* mutated patients. The efficacy of IFN-I therapy would depend on whether it is administered early in the course of the disease. Patients with a severe course of COVID-19 are usually admitted to the hospital after a few days at home making it difficult to identify those in need of IFN-I treatment. Indeed, inappropriate administration of IFN-I to the wrong patients or at the wrong time point could be counterproductive by triggering the cytokine storm. A more attractive therapeutic option would be IFNγ, which is not only useful in patients with hypomorphic mutations but, in addition, can stabilize the inflammatory response and does not require timely administration (Mantovani S. et al., 2022).

5.4.2 Autosomal dominant COVID-19 due to *ADAMTS13* ultra-rare variants

ADAMTS13 gene (localized on chromosome 9) encodes for a disintegrin-like and metalloprotease with thrombospondin type 1 motif, 13 which is a plasma glycoprotein with protease activity that plays a fundamental role in platelet adhesion and aggregation on vascular lesions. Congenital Thrombotic Thrombocytopenic Purpura (cTTP) is a severe autosomal recessive disorder due to *ADAMTS13* rare variants and is characterized by uncleaved ultra-large vWF and thrombotic microangiopathy, frequently triggered by infections. Common polymorphisms in the same enzyme are reported to be involved in thrombosis (Pagliari M.T. et al., 2021; Haybar H. et al., 2018). Moreover, ADAMTS13 protein production is positively induced by oestrogen, and this reflects the greater penetrance of acquired or congenital TTP in middle aged females (over 50 years), whose oestrogen levels start to decrease in relation to males (Powazniak Y et al., 2011). It is well recognized that coagulation abnormalities with an increased risk of thrombosis are one of the complications of severe COVID-19 disease, characterized by a high level of IL-6 and D-

dimer often together with a reduction in platelets. The reduced activity of ADAMTS13 is already reported to be associated with a severe COVID-19 outcome (Hafez W et al., 2022).

Exome analysis of 2988 SARS-CoV-2-infected subjects of different severities, belonging to GEN-COVID cohort, stratified by sex and adjusted by age and applying an OLR Model, shows an association between *ADAMTS13* ultra-rare variants and severity in female patients with an OR = 3.32. No significant association was found in male patients (p-value = 0.252) (Zguro K et al., 2022 b).

Most of the subjects (106/124) with heterozygous ultra-rare variants had severe COVID-19 disease requiring hospitalization (85.5%). The remaining 18 subjects (14.5%) were not-hospitalized patients. The majority of not-hospitalized patients (14 subjects) were either females under 50 years or males over 50 years of age. For the females under 50, a protective role of oestrogen, which increases *ADAMTS13* transcript, can be envisaged (**Figure 13**). Among the hospitalized patients, there were also three females younger than 10 months, (Zguro K. et al., 2022 b).

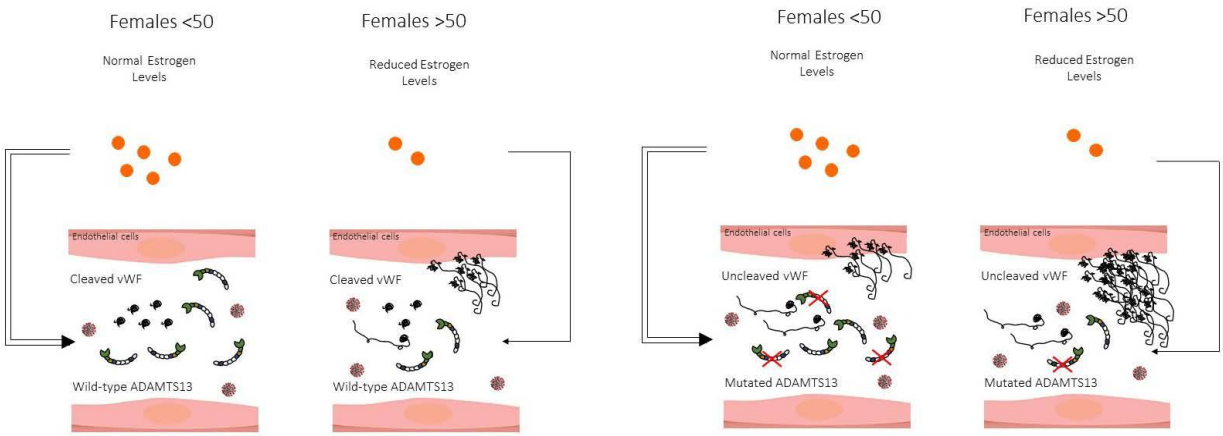


Figure 13. ***ADAMTS13*** **ultra-rare variants and severity in female patients.** In women over 50 years of age with heterozygous variants in ADAMTS13 there is a greater formation of VWF multimers and therefore a greater thrombotic risk, due both to the lower enzymatic activity of ADAMTS13 and to the reduction of the protective effect of oestrogens. From Zguro K. et al., 2022 b.

In addition, for eleven subjects (6 cases with *ADAMTS13* variants and 5 controls without variants) it was possible to perform an ad hoc blood drawn in order to assess ADAMTS13 activity after SARS-CoV-2 infection. Carriers of ultra-rare variants show a significant reduction of ADAMTS13 activity, p-value = 0.017 (Wilcoxon test), as expected for heterozygous subjects (**Figure 14**).

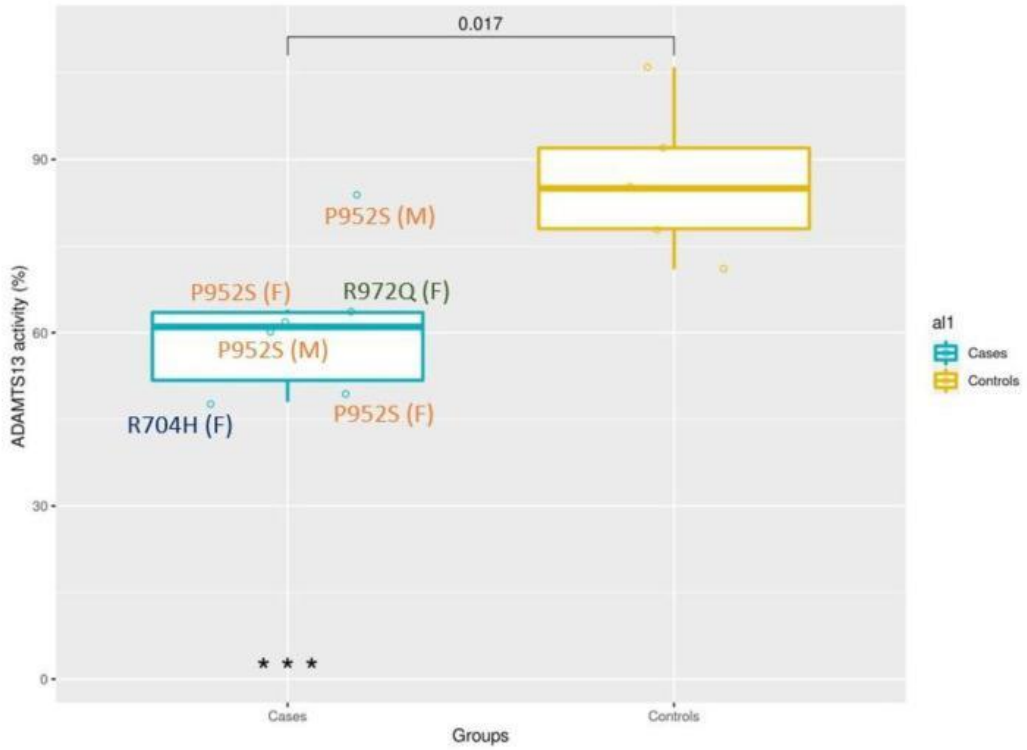
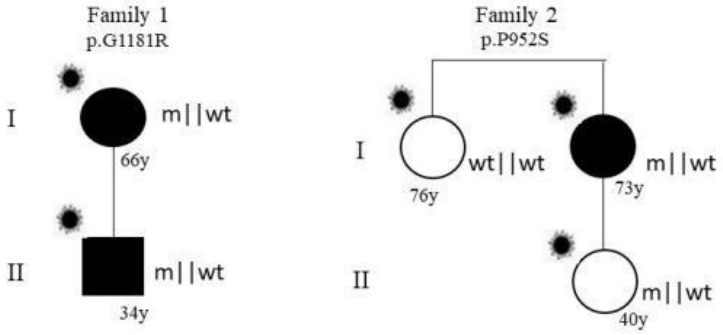


Figure 14. Heterozygous *ADAMTS13* ultra-rare variants are related to a reduction of protein detection. Box plot of patients with one ultra-rare variant (6 cases) and patients without ultra-rare variants (5 controls). The presence of ultra-rare variants is associated with a reduction of ADAMTS13 activity (p-value = 0.017 at Mann–Whitney U test). From Zguro K. et al., 2022 b.

During hospitalization, both males and females with heterozygous *ADAMTS13* variants have a tendency for hyper-inflammation (CRP mean 39, p = 0.005), higher D-dimer (mean 3024, p = 0.03), platelets consumption (platelet count mean 180, p = 0.07) and haemolysis (LDH mean 444, p = 0.009). The correlation is sustained mainly by females ≥50 years (CRP mean 55, p = 0.005; LDH mean 506, p = 0.006933) and males <50 years (platelet mean 153, p = 0.052).

Complete clinical and molecular data were available for two families (Figure 15). Segregation analysis demonstrates that the disorder segregates as an autosomal dominant disease conditioned by SARS-CoV-2 infection, sex, and age. In the first family, the 66-year-old female who required oxygen support (Clinical category 2) transmitted the variant to the 34-year-old son who required CPAP treatment (Clinical category 3). In the second family, the 73-year-old female treated by oxygen support (Clinical category 2) transmitted the variant to the 40-year-old daughter who was oligo-symptomatic (Clinical category 0), likely due to the relatively young age; her sister, the 76-year-old without the variant, was oligo-symptomatic (Clinical category 0), (Zguro K. et al., 2022 b) (**Figure 15**).



Family ID	Pedigree position	Gender	Age	Clinical Category	Signs/Symptoms	Infected	*ADAMTS13* p.G1181R
1	I-1	F	66	2	Bilateral pneumonia, oxygen treatment	+	Heterozygous
1	II-1	M	34	3	Bilateral diffused pneumonia, CPAP treatment	+	Heterozygous

Family ID	Pedigree position	Gender	Age	Clinical Category	Signs/Symptoms	Infected	*ADAMTS13* p.P952S
2	I-1	F	76	0	Oligosymptomatic, not hospitalized	+	Wild type
2	I-2	F	73	2	Bilateral diffused pneumonia, oxygen treatment	+	Heterozygous
2	II-1	F	40	0	Oligosymptomatic, not hospitalized	+	Heterozygous

Figure 15. Segregation analysis. Pedigree (upper panel) and respective segregation of *ADAMTS13* variant and COVID-19 status (lower panel) are shown. Squares represent male family members; circles represent females. A virus cartoon close to the individual symbol indicates individuals infected by SARS-CoV-2. The inheritance pattern appears that of an autosomal dominant disorder conditioned by SARS-CoV-2 infection, sex, and age. From Zguro K. et al, 2022 b.

These data confirm that ultra-rare variants in a heterozygous state lead to a rare form of COVID-19 characterized by hyper-inflammation signs, which segregates in families as an autosomal dominant disorder conditioned by SARS-CoV-2 infection, sex, and age. This has clinical relevance due to the availability of drugs such as Caplacizumab, which inhibits vWF–platelet interaction, and Crizanlizumab, which, by inhibiting P-selectin binding to its ligands, prevents leukocyte recruitment and platelet aggregation at the site of vascular damage. These two drugs are likely to replace the reduced activity of the metalloproteinase due to certain mutations and therefore they could also be useful in decreasing hyper-inflammation signs in heterozygous *ADAMTS13* patients (Zguro K. et al., 2022 b).

5.4.3 *CFTR* gene and severe COVID-19

The *CFTR* gene, localized on chromosome 7, encodes an ATP-binding cassette (ABC) transporter that functions as a low conductance Cl(-)-selective channel gated by cycles of ATP binding and hydrolysis at its nucleotide-binding domains (NBDs) and regulated tightly by an intrinsically disordered protein segment distinguished by multiple

consensus phosphorylation sites termed the regulatory domain. This gene is expressed on the apical membrane of epithelial cells, mainly in lung, pancreas, liver and intestine (Elborn J.S. et al., 2016), where also *ACE2* (angiotensin-converting enzyme 2), the entry receptor for SARS-CoV-2, is localized (Hoffmann M. et al., 2020).

When both *CFTR* alleles are mutated, in a homozygous or compound heterozygous state, and the global CFTR activity is impaired by more than 70% (and often more than 95%), patients are affected by cystic fibrosis (CF). CF is a multiorgan disease characterized by high viscosity of secreted fluids, causing plugs and obstructions, and by an abnormal inflammatory response, independent but aggravated by infections, leading to respiratory failure and premature death (Roesch E.A. et al, 2018; Elborn J.S. et al., 2016).

It is known that heterozygotes for *CFTR* variants (CF carriers) have a reduction in CFTR expression and function, based on the type of the pathogenic variant. Therefore, they have high risk of developing CF-related conditions (Miller A.C. et al., 2020). In particular, they are significantly more susceptible to airway and sinus infections, pneumonia, pancreatic injury and hepatitis (Miller A.C. et al, 2020; Polgreen P.M. et al., 2018). These clinical problems are already all described in the severe form of COVID-19 (Wiersinga W.J. et al., 2020; Mahmudpour M. et al., 2020; Wang Y. et al., 2020).

The ML LR LASSO model extracted the *CFTR* gene, in the presence of ultra-rare variants, as one of the genes associated with severe form COVID-19 for both sexes (Fallerini C. et al., 2022).

On a cohort study of 874 Italian individuals diagnosed with COVID-19 and recruited through the GEN-COVID multicenter study (NCT04549831), during the first pandemic wave in Italy, it has been identified those that are carriers of single pathogenic variants of the *CFTR* gene and evaluate their clinical course, in order to determine to what extent CFTR impairment contributes to COVID-19 susceptibility and severity (Baldassarri M. et al., 2021).

Seventeen CF-causing variants were identified in 41 COVID-19 patients. Specifically, 13 patients harboured the genomic deletion of three base pairs resulting in the loss of phenylalanine at amino acid position 508 of the CFTR protein (p.Phe508del). Nobody was carrier of the TG12-5T polymorphism, nor the TG13-5T known to reduce CFTR function. Forty patients (4.58% of the whole cohort), 26 males (65%) and 14 females (35%), were identified as carriers of one CF-causing variant in the *CFTR* gene. Whereas, one patient (male, 52 years) who was found to have two pathogenic variants, was excluded (Baldassarri M. et al., 2021).

Specifically, hospitalized CF carriers develop a form of COVID-19 more likely characterized by acute respiratory distress syndrome, high inflammatory response, and, in some cases, hyper-lipasemia. Moreover, carriers who have needed invasive mechanical ventilation are significantly younger (mean age of 51 years) than non-carriers in the same clinical category and the majority (83.33%) are males. Therefore, those evidences indicate that CF carriers may be prone also to develop a severe COVID-19, and even at a younger age compared to non-carriers. Furthermore, the higher prevalence of male individuals confirms a world trend that defines the male sex as a relevant risk factor for severe COVID-19 (Baldassarri M. et al., 2021).

Regarding risk factors for mortality, while age ≥ 75 and AST ≥ 40 U/L are relevant at all the studied time points and at 28/60-days respectively, the status of CF carrier, LDH ≥ 400 and age ≥ 75 are determinants of mortality at day 14. In particular, after adjustments for age, sex and comorbidities, being a carrier of known CF-causing variants appears to be a relevant factor (HR, 3.10, CI, 1.09–8.85) determining early mortality. However, it is important to underline that at the later time-points (day-28 and day-60) CF carrier status does not appear to be anymore a risk factor for death. These data reveal that COVID-19 mortality is linked to time-dependent factors and that CFTR-related early events like cytokine storm may be accountable for early death (Baldassarri M. et al., 2021), (**Figure 16**).

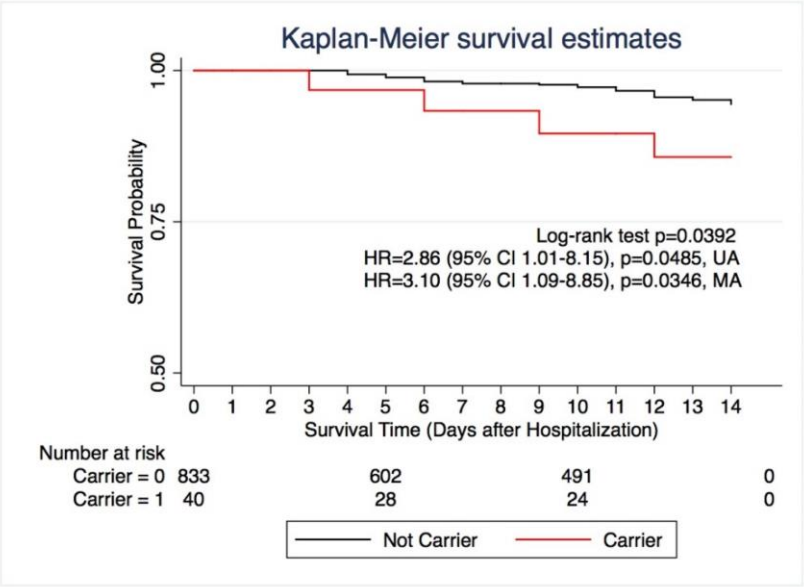


Figure 16: **Survival Study: Kaplan–Meier 14-day survival study comparing carriers of single CF causing variants (red) and noncarriers (black).** Log-rank test p = 0.0392; Univariable Cox analysis (UA) HR = 2.86; 95% CI 1.01–8.15; p = 0.0485. Multivariable Cox analysis (MA) with adjustment for confounding variables HR = 3.10; 95% CI, 1.09–8.85; p = 0.0346. From Baldassarri M. et al, 2021.

Hence, the status of CF carrier, given its high prevalence, should be investigated in COVID-19 hospitalized patients in order to identify subjects that, being at risk of severe disease, would benefit of intensive surveillance and personalized therapy.

In addition, applying the ML LR LASSO model to a large international cohort of COVID-19 patients, the *CFTR* gene was extracted as an important factor involved in the modulation of COVID-19 outcomes (Fallerini C. et al., 2022).

On the contrary, *CFTR* ultra-rare alleles have loss-of-function mechanisms and are associated with severity in both older women and men (age ≥ 50). Overall, these alleles led to global residual activity of CFTR between 50 to 79%. The reduced function leading to severe COVID-19 could be achieved by having one cis complex allele (dominant disorder) or two hypomorphic alleles in trans (recessive disorder). In the first case, severe COVID-19 would be transmitted as an autosomal dominant disorder, while in the second case, an autosomal recessive disorder

The rare complex allele [G576V;R668C] is associated with a milder disease through a gain-of-function mechanism with increased activity to 110%, especially in younger women (age < 50 years) and older men (age ≥ 50 years). On the contrary, *CFTR* ultra-rare alleles have LOF mechanisms and are associated with a severe disease in both older women and men (age ≥ 50). Overall, these alleles cause a global residual activity of CFTR between 50 to 79%. The reduced function causing severe COVID-19 could be achieved by having one cis complex allele (dominant disorder) or two hypo-morphic alleles in trans (recessive disorder). In the first case, severe COVID-19 would be transmitted as an autosomal dominant disorder, while in the second case, an autosomal recessive disorder (Baldassarri M et al., 2022).

The distribution of the gain-of-function and loss-of-function *CFTR* alleles among the COVID-19 clinical categories showed an age- and sex-dependent pattern. In particular, younger females and older males with mild symptoms shared more frequently the gain-of-function complex alleles, while older men and women with the most severe disease shared the loss-of-function ones. This finding may indicate a putative synergic effect of oestrogens and *CFTR* variants in modulating the response to SARS-CoV-2 infection. Thus, *CFTR* gain-of-function variants may either counterbalance the inhibitory effects of oestrogens or have a synergic activatory effect, therefore leading to the manifestation of mild COVID-19 disease in younger females and older males, known to have higher oestrogen levels compared to post-menopausal women (Baldassarri M et al., 2022).

Concerning the significant prevalence of loss-of-function *CFTR* alleles/variants in the most severe clinical categories, these data confirm what previous work has already reported (Baldassarri M. et al., 2021) showing that carriers of single pathogenic *CFTR* variants are more likely to undergo severe COVID-19 with a high risk of 14-day mortality. Oestrogen levels and the reported reduced *CFTR* expression as a consequence of aging may further contribute to the severity of COVID-19 disease in older men and women (Baldassarri M et al., 2022).

Concerning the only one individual affected by CF, a 51-year-old male hospitalized with low-flow oxygen as respiratory support, this is in line with other reports showing that CF patients undergo mild COVID-19 (Mathew H.R. et al., 2021).

Actually, CF patients may have several potential protective factors: (i) the treatment with CF modulators that reduces the odds of hospitalization (Carr, S.B.;2021); (ii) their habit to social distancing and infection control practices; (iii) the impaired membrane expression of *CFTR* that would not allow for binding with the SARS-CoV-2 spike protein (Caohuy, H. et al. 2022); (iv) their altered signalling of ACE and ACE2 (Bezerri et al., 2021); (v) the use of certain drugs, such as azithromycin, that may protect against infections (Mathew H.R. et al., 2021); and (vi) the systemic levels of ATP intrinsically elevated in CF patients (Caracciolo et al. 2021, Baldassarri M et al., 2022).

5.5 COVID-19: a polygenic disease.

As already stated, the Mendelian monogenic model clarifies only a small fraction of severe COVID-19 cases. In particular, this model is caused by rare or ultra-rare variants which, although they have a great impact on the protein, are only found in a minority of cases. In addition, COVID-19 is a complex and multifactorial disease, characterized by high clinical variability, with possible multi-organ involvement in the most severe forms.

For this reason, it is expected that, in most cases, thousands of genes are involved to determine together the severe phenotype. As already described, the model allowed to identify over 7,200 variants associated with the different severity of the phenotype.

These variants include, in addition to rare variants, also common polymorphisms. Common polymorphisms are present in more than 5% of the general population and for this reason their protein functional impact is expected to be lower than that of rare variants. However, the combination of multiple common low-impact variants can lead to a severe form of COVID-19.

The method presented in this study allowed to identify some genes with common functional polymorphisms that may contribute to the variability of the clinical manifestation.

5.5.1 Severe COVID-19 in males due to hyper-inflammatory response: the role of *AR* gene and *CYP19A1* gene.

The gene encoding androgen receptor (AR), alternatively known as the dihydrotestosterone receptor, is located on the X chromosome. It is mutant in the androgen insensitivity syndrome (AIS), formerly known as the testicular feminization syndrome (TFM), and in Kennedy spinal and bulbar muscular atrophy (SBMA).

AR is a ligand-dependent transcription factor and belongs to the nuclear receptor family. These receptors recognize canonical androgen response elements (AREs), which are inverted repeats of 5-prime-TGTTCT-3-prime. The major domains of AR include N- and C-terminal activation domains, which are designated activation function-1 (AF-1) and AF-2, a ligand-binding domain, and a polyglutamine tract (Callewaert L. et al., 2003). The polymorphic polyQ tract ranges between 9 and 36 repeated CAG units in the normal population. In the full size AR, the deletion of the polyglutamine tract results in an increase in the transactivation through canonical AREs. In vitro and in vivo studies have demonstrated that the transactivation potential of AR is inversely correlated to repeat length, and Q-tract size can significantly influence androgen dependent physiological functions (Callewaert L. et al., 2003). AR polyQ length correlates with receptor functionality, with shorter polymorphic glutamine repeats typically associated with higher and longer PolyQ tracts with lower receptor activity (Callewaert L. et al., 2003). AR is expressed in both males and females, but the bioavailability of its ligands T and dihydroT (DHT) differs significantly, being much higher in males.

Many evidence prove the concept that androgens are relevant to both SARS-CoV-2 infection and COVID-19 disease presentation; however, they seem to have a Janus bifacial way of action (Wambier CG et al., 2020; Pozzilli P et al., 2020). On one side, androgens promote the transcription of the *TMPRSS2* gene that encodes a serine protease known to prime the spike (S) protein of coronaviruses, facilitating viral entry into the cells (Hoffmann M et al., 2020). On the other hand, hypogonadism is known to correlate with severe COVID-19 (Rastrelli G et al., 2020) and other chronic conditions, partly due to the loss of attenuation of the inflammatory immune response exerted by testosterone (T).

The ML LR LASSO model extracted the *AR* gene, in the presence of polyQ repeats<23, as one of the genes associated with a protective role to severe COVID-19 for males (Fallerini C. et al., 2022).

In order to test the role of common poly-amino acid repeat polymorphisms in determining COVID-19 clinical severity, a nested case-control study (NCC) was performed, selecting the extreme phenotypic ends of the Italian GEN-COVID cohort (Daga S et al., 2021). Cases were selected according to the following inclusion criteria: i. CPAP/biPAP ventilation (230 subjects) (Clinical category 3); ii. endo-tracheal intubation (108 subjects) (Clinical category 4). As controls, 300 subjects were selected using the sole criterion of not requiring hospitalization (Clinical category 0). A similar Spanish cohort, composed of male COVID-19 patients (117 cases and 41 controls) was used to validate the results in another representative European population highly impacted by COVID-19 (Baldassarri M. et al., 2021 b).

In order to validate the results on AR obtained by LASSO logistic regression, the number of triplets was sized in the male subset (351 subjects) using the gold standard technique that uses a fluorescent PCR reaction followed by the use of GeneScan Analysis software® (Applied Biosystems). Based on the *AR* polyQ length, male patients were subdivided into two categories, those having a number of PolyQ repeats less than or equal to 22 repeats, and those having a number of PolyQ repeats greater than or equal to 23 repeats, being 23 repeats the reference sequence on genome browsers and the reported cut-off value (Baldassarri M. et al., 2021 b).

It was found that PolyQ repeats below 22 are enriched in the asymptomatic cohort of males. The difference was statistically significant in the group of males younger than 60 years of age in which genetic factors are expected to have a major impact (p-value 0.024 by x2 test). The association with shorter repeats (≤ 22) and protection was confirmed (p-value 0.014 by x2 test) in the Spanish cohort (Baldassarri M. et al., 2021 b).

To functionally link the length of the PolyQ repeats to AR functionality, TT in 183 men was measured using LCMS/MS. TT was higher in patients carrying ≥ 23 vs ≤ 22 glutamines, suggestive of impaired negative feedback at the level of the hypothalamus and pituitary gland. (Baldassarri M. et al., 2021 b), (**Figure 17**).

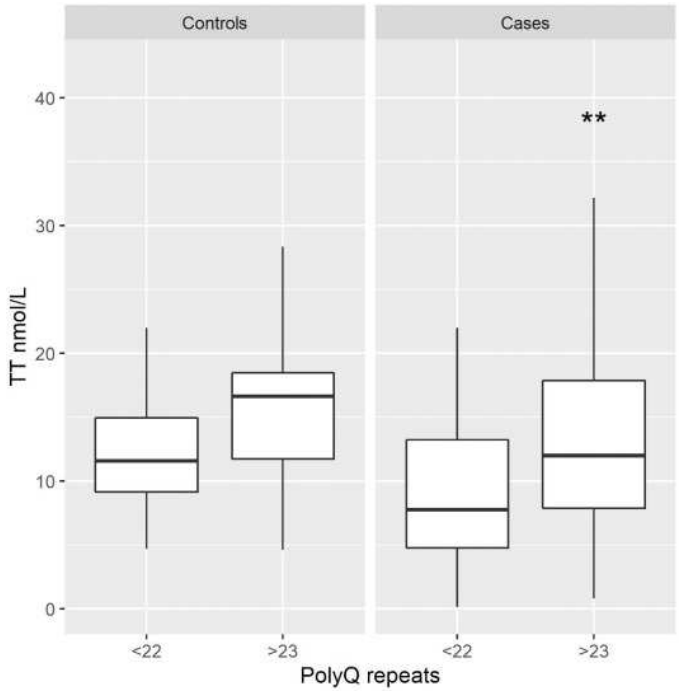


Figure 17. Relationship between Total Testosterone and polyQ repeats in the case and the control group. Box-plot showing values of Total Testosterone (TT), expressed in nmol/L, in subjects with shorter (≤ 22) and longer (≥ 23) polyQ repeats in *AR* gene grouped between controls (left panel) and cases (right panel). The TT median value, represented by the black horizontal line, is higher in patients with ≥ 23 polyQ repeats in the case group, (**p-value = 0.023; Mann-Whitney U test). No statistically significant difference was present in the control group (p-value = 0.088; Mann-Whitney U test). From Baldassarri M. et al., 2021 b.

While this is compensated in healthy subjects, during non-gonadal illnesses (NGI) such as COVID-19, some patients are unable to compensate for the reduced AR activity with higher T levels. The result is a status of reduced androgenicity even in the presence of apparently normal T values (Zitzmann M. et al., 2003).

To evaluate whether the AR receptor reduced activity resulted in signs and symptoms of hypogonadism, subjects were interviewed, post-infection, using a modified version of the Androstest® (Millar A.C. et al., 2016). Interviews were available for 61 subjects (43 short and 18 long) representative of the extremes genotypes (≤ 19 and ≥ 25 repeats) of the cohort. The results indicate a trend toward clinical hypogonadism for those with longer repeats.

As T is known to have an immunomodulatory activity attenuating inflammatory immune responses (Eendebak R.J. et al., 2016; Zitzmann M. et al., 2003; Gubbels Bupp M.R. et al., 2018; Hoebe K. et al., 2004; Janeway Jr C.A. et al., 2002; Lai J.J. et al., 2012; Medzhitov R. et al., 2000), it was hypothesized that a long PolyQ repeat would lead to a pro-inflammatory status heralded by increased proinflammatory markers (Pierotti S. et al., 2010; Mohamad N.V. et al., 2019) by conferring decreased AR transcriptional activity.

Conversely, men with a more active receptor (short PolyQ tract) would be protected because they can tame the inflammatory response and increase survival regardless of serum T levels. In particular, CRP, one of the main inflammatory markers, was higher in subjects with a long *AR* PolyQ tract. This observation not only is in line with the known anti-inflammatory function of T, but also reinforces the functional importance of the *AR* PolyQ tract and its association with COVID-19 clinical outcome. Furthermore, this observation suggests that CRP is hierarchically more relevant than serum T level, which can be inappropriately normal and mask a status of low androgenicity in men with a long PolyQ repeat (Baldassarri M. et al., 2021 b).

Overall, these results confirmed the already known association between hypogonadism or long poly-Q trait and severe forms of COVID-19 (Rastrelli G. et al., 2020; McCoy J. et al., 2020) and other chronic obstructive pulmonary diseases (Van Vliet M. et al., 2005; Mohan S.S. et al, 2015).

Ultimately, this the first genetic polymorphism predisposing some men to develop a more severe disease irrespectively of age. For this reason, sizing the *AR* poly-glutamine repeat has important implications in the diagnostic pipeline of patients affected by life-threatening COVID-19 infection. In addition, the potential use of testosterone could be hypothesized as an add-on therapy for patients with severe COVID-19 for a defective androgen signalling, defined as ≥ 23 PolyQ repeats in the AR gene, and inappropriately low levels of circulating androgens.

More recently, through International collaboration, combining human and animal data, CYP19A1 (the testosterone to estradiol converting enzyme) gene has been identified as a crucial factor of severe disease outcome upon SARS-CoV-2 infection in males. In particular, the Thr201Met variant in *CYP19A1*, which increases CYP19A1 enzyme (aromatase) activity in general, is expected to contribute to severe COVID-19 in up to 6.2% of all hospitalized male and female patients. Remarkably, in male patients, the Thr201Met variant showed a calculated penetrance under a monogenic model of 68.7%. Men with this specific variant were more likely to be hospitalized with severe COVID-19 compared to men without this variant or compared to women with COVID-19. Moreover, elevated CYP19A1 expression was still prominent in the lungs of male COVID-19 patients at the time of death, when in most cases viral RNA was no longer detectable.

In the golden hamster model, SARS-CoV-2 infection causes increased *CYP19A1* expression in the lung that is associated with dysregulated plasma sex hormone levels and reduced long-term pulmonary function in males but not females. In particular, male and

female hamsters infected with SARS-CoV-2 presented comparable weight loss until day 6 post infection (the acute phase of infection) but showed significant differences from day 7 until 14 post infection (recovery phase). SARS-CoV-2-infected male hamsters presented delayed weight gain and impaired lung function even 1 week after recovery from SARS-CoV-2 infection.

Treatment of SARS-CoV-2-infected hamsters with a clinically approved CYP19A1 inhibitor (letrozole) improves impaired lung function and supports recovery of imbalanced sex hormones specifically in males (**Figure 18)** (Stanelle-Bertram et al., 2023). Treatment of SARS-CoV-2-infected animals with the clinically approved CYP19A1 inhibitor letrozole improved long-term lung health in males, suggesting letrozole as a promising drug candidate for further assessment regarding individualized therapy in humans (Stanelle-Bertram et al., 2023).

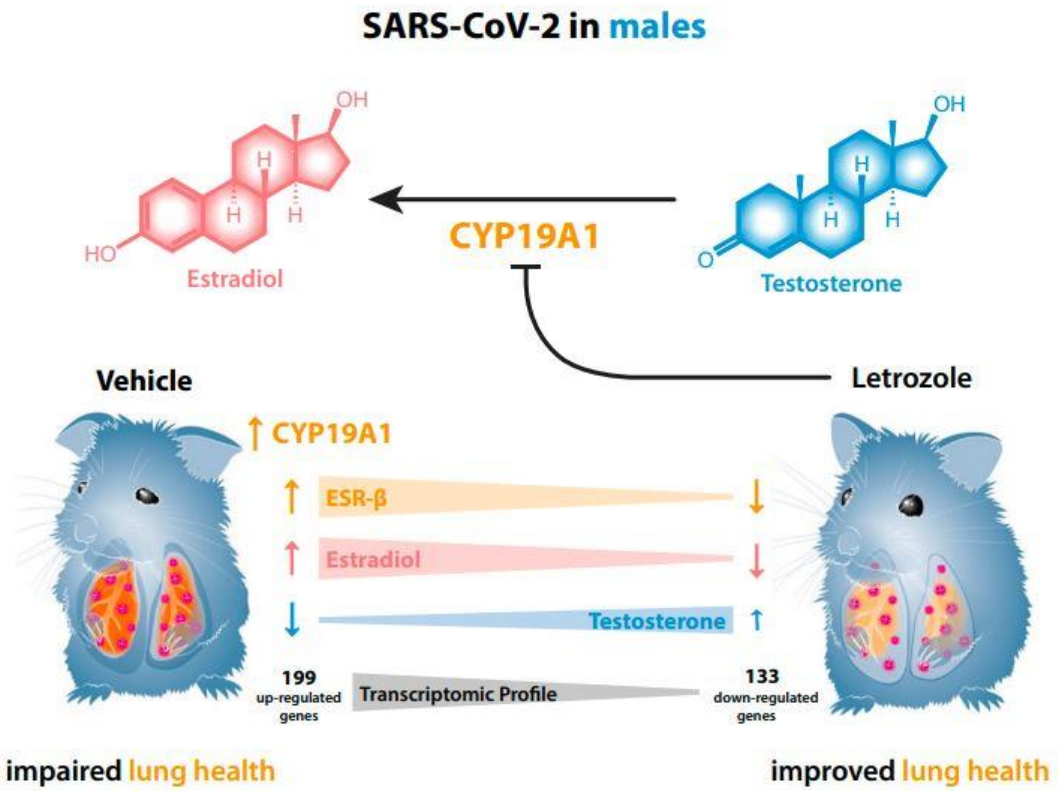


Figure 18. Treatment of SARS-CoV-2-infected hamsters with the CYP19A1 inhibitor letrozole. Treatment of hamsters with the CYP19A1 inhibitor letrozole results in reduced estradiol and ESR-ß levels, while testosterone levels are elevated. Moreover, letrozole treatment results in the downregulation of 133 genes in SARS-CoV-2 infected males, as compared to the SARS-CoV-2 infected, vehicle-treated control group. From Stanelle-Bertram et al., 2023.

5.5.2 The polymorphism L412F in *TLR3* inhibits autophagy and is a marker of severe COVID-19 in males

The *TLR3* gene, localized on chromosome 4, encodes an endosomal transmembrane receptor that recognizes double-stranded RNA (dsRNA) and activates signalling cascades that culminate in innate antiviral immunity against a number of viral infections. TLR3 belongs to the family of Toll-like receptors (TLRs) that are pattern recognition receptors expressed by various types of cells. Activation particularly induces production of interferons, such as IFNA1, IFNB1, and IFNG), as well as inflammatory cytokines, such as IL6 (Mukherjee S. et al., 2019; Perales-Linares R. et al., 2013; Totura A.L. et al., 2015). After the viral infection, the primary response genes induced by this activation were co-regulated by the NFKB pathway, and the IRF3 pathway, which play an essential role in the immune response. This results in the production of various cytokines, including TNF (tumour necrosis factor), activating immune responses. Additional secondary response genes were activated by autocrine and paracrine secretion of IFNB. However, increased inflammatory responses can lead to a higher risk for pneumonia and autoimmune diseases. Accordingly, a protective effect against fatal pneumonia has been reported in the absence of TLR3 (Matsumoto M. et al., 2011; Schulz O. et al., 2005; Suresh M.V. et al., 2019).

The ML LR LASSO model extracted the *TLR3* gene, in the presence of the p.Leu412Phe (L412F) polymorphism (rs3775291), both in heterozygous and homozygous state, as a severity marker for COVID-19 both in males and females (Fallerini C. et al., 2022).

The L412F polymorphism has an overall allele frequency of about 20%, ranging from 30% in European to 0.88% in African (mainly sub Saharan) populations (GnomAD browser). This functional polymorphism is known to decrease TLR3 expression on the cell surface (Ranjith-Kumar C.T. et al., 2007). Moreover, it leads to poor recognition of SARS-CoV-2 dsRNA, during replication, and has been recently associated with SARS- CoV-2 susceptibility and mortality (Teimouri H. et al., 2020). TLR3, as with other TLRs, determines susceptibility to infections through autophagy (Franco L.H. et al., 2017), which is one of the major cell defense mechanisms against pathogens (Levine B. et al., 2011).

A role for autophagy is reported in SARS-CoV-2 infection. In particular, SARS-CoV-2 can inhibit autophagy resulting in accumulation of autophagosomes and inhibition of viral clearance that, together with immune dysfunction and the activation of numerous

inflammatory cytokines, leads to a more severe form of COVID-19 (Carvalho-Schneider C. et al., 2021; Hoffmann M. et al., 2020; Miao Y. et al., 2020).

In order to further investigate the mechanisms underlying the diverse susceptibility to COVID-19, a nested-control study was performed within the Italian GEN-COVID cohort: it has been confirmed the role of *TLR3* L412F polymorphism in susceptibility to SARS-CoV-2 and further defined the potential mechanisms by which this effect is exerted (Croci S. et al., 2022). Individuals showing severe COVID-19 (cases) and those with no sign of the disease (controls) were compared; patients were divided into two categories, those having the poly-morphism in heterozygous or homozygous state and those homozygous for the WT allele. The prevalence of L412F polymorphism was significantly higher in cases compared to controls (p-value 2.8 × 10−2). The global allele frequency of L412F in the study cohort (cases and controls) was 29.38%, comparable to the allele frequency of 29.79% reported in the European (non-Finnish) population (https://gnomad.broadinstitute.org/) (Croci S. et al., 2022).

In addition, stratifying the study cohort by gender, the statistically significant difference increased in the sub-cohort of males giving an Odds Ratio of 1.94 (95% confidence interval, 1.23 to 3.06; p = 3.8x10−3), whereas it was lost in the sub-cohort of females (p-value 5.8 × 10−1). Then, it was investigated the prevalence of patients carrying L412F in heterozygous or homozygous states in all 4 categories of COVID-19 clinical severity, considering only male subjects regardless of age (n = 665) and it was found that the prevalence of carriers directly increased with the severity of COVID-19 (Croci S. et al., 2022), (**Figure 19**).

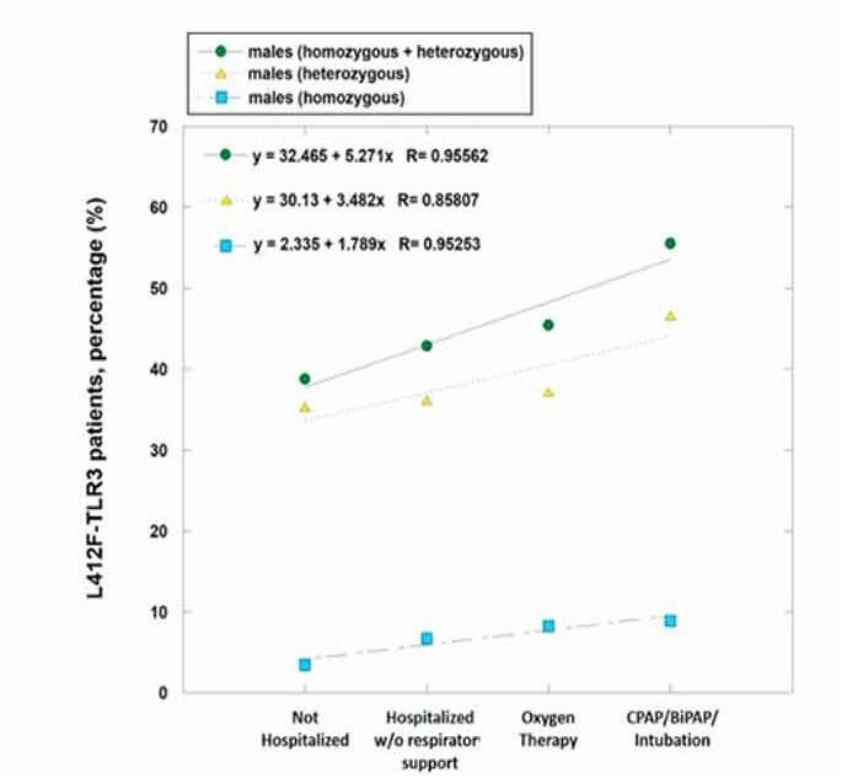


Figure 19. **Distribution of carriers of the polymorphism L412F in homozygous or heterozygous states stratified by clinical category.** Modified from Croci S. et al., 2022.

In order to investigate how this polymorphism impacted receptor function, we used HEK293 cell lines transfected with both versions of the polymorphism. In HEK cells transfected with TLR3_L412F, the number of autophagosomes (APs) was reduced by poly(I:C) stimulation (a specific TLR3 agonist), demonstrating a block in AP synthesis and a reduced autophagic flux. A statistically significant reduced survival at 28 days was shown in L412F COVID-19 patients treated with the autophagy-inhibitor hydroxychloroquine (p = 0.038) (Croci S. et al., 2022), (**Figure 20**).

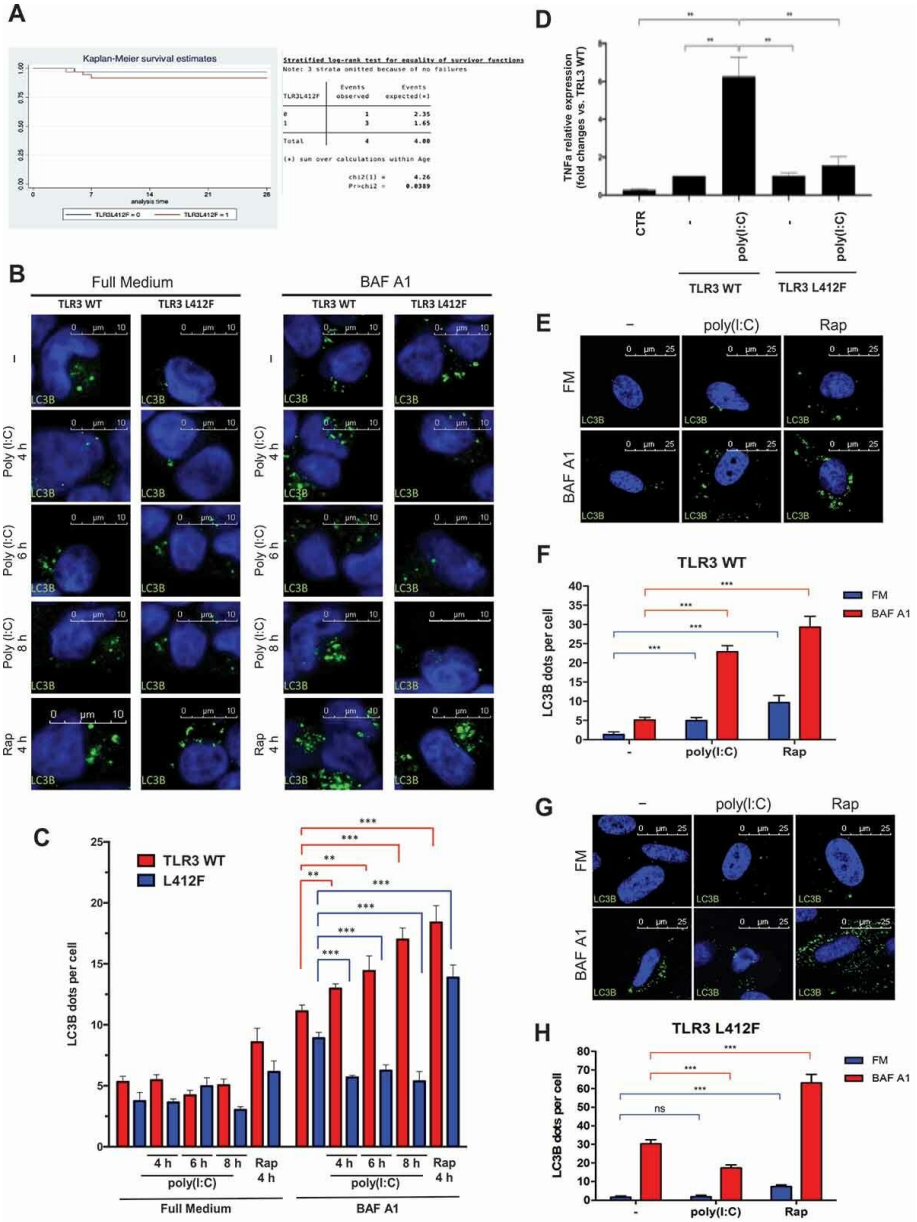


Figure 20. Analysis of autophagy in *TLR3*_L412F-expressing cells. (A) 28-day survival study of *TLR3*-L412F carriers vs not-carriers in the group treated with hydroxychloroquine. (B - C) Analysis of autophagy in HEK- KO cells expressing wild type or L412F mutant proteins. (D) Analysis of TNF mRNA expression in HEK-KO cells expressing wild type or L412F mutant proteins. (E - F) Analysis of autophagy in normal human

fibroblasts (NDHF) from subjects expressing the TLR3_WT (G and H) same as in E-F, but fibroblasts are homozygous for the *TLR3*_L412F receptor. Statistical analysis was performed using Student's t test. Means § SEM for each value are shown in the graphs. ns = not significant; ** = $p < 0.01$; *** = $p < 0.001. < 0.001$. From Croci S. et al., 2022.

It is known that *TLR3* variant L412F has been associated with a wide range of autoimmune diseases including Addison disease and hypothyroidism. Alteration of autophagic processes causes the onset of autoimmunity due to increased survival and reduced apoptosis of self-reactive lymphocytes (Keller at al., 2017). HLA has been shown to be implicated in disease severity and clinical outcome of patients with COVID-19. An increased frequency of autoimmune disorders such as co-morbidity was found in L412F COVID-19 males with specific class II HLA haplotypes, DR3/DQ2, prone to autoantigen presentation, especially diabetes (25% of cases). These results suggest that the combination of L412F in TLR3 and a specific class II HLA haplotype puts male patients at risk of post-COVID autoimmune exacerbation emphasizing the need for appropriate follow-up (Croci S. et al., 2022).

In conclusion, the *TLR3* L412F polymorphism makes males, in whom after puberty testosterone lowers *TLR3* expression, at risk of severe COVID-19 in a context of a polygenic model. In addition, based on the impairment of autophagy, it would be important to reinterpret clinical trials with patients stratifying HCQ according to L412F and to suggest, instead, treatment with IFN (in particular IFN-γ) avoiding HCQ. Lastly, the combination of L412F in *TLR3* and specific HLA class II haplotypes may put male patients at risk of post-acute sequelae of SARS-CoV-2 infection pointing to the need for an appropriate follow-up (Croci S. et al., 2022).

5.5.3 *SELP* Asp603Asn and severe thrombosis in COVID-19 males

The P-selectin (*SELP*) gene, localized in chromosome 1, encodes an adhesion molecule, expressed at the surface of activated cells, that mediates the interaction of activated endothelial cells or platelets with leukocytes, playing a key role in thrombosis (Blann A.D. et al., 2003; Merten M. et al.; 2004).

As already reported, it is now widely known that COVID-19 is a systemic disease, characterized by dysregulation of the immune system and by a hypercoagulable state (Tang N. et al., 2020). It is not yet clear which is the mechanism behind this increased susceptibility, but it may certainly be partly due to genetic factors of the host.

The ML LR LASSO model extracted the *SELP* gene, in the presence of the polymorphism c.1807G>A, p.D603N (rs6127) in homozygous state, as associated with severe COVID-19 in males (**Figure 21 a**), (Fallerini C. et al., 2022).

This polymorphism has already been described as a functional polymorphism associated with thrombotic risk in various conditions (Ay C. et al., 2008; Tregouet D.A. et al., 2002). Indeed, it has been shown that the polymorphism affects the binding of P-selectin to its ligand on leukocytes, possibly making the protein more efficient at recruiting leukocytes to the endothelium (Tregouet D.A. et al., 2002).

Furthermore, significantly increased P-selectin and other pro-thrombotic biomarkers concentration in plasma samples of severe COVID-19 patients compared to healthy controls has been recently reported (Bongiovanni D. et al., 2021; Manne B.K. et al, 2020).

In a sub-cohort of 513 male subjects the homozygosity of the polymorphism c.1807G>A, p.D603N (rs6127) is found to be associated with severity (odds ratio=2.27, 95% Confidence Interval 1.54–3.36). As the *SELP* transcription is inhibited by androgens (Karolczak N. et al., 2018), the strength of the association should increase with age. Interestingly, the OR (2.42) in males aged ≥ 50 years with respect to the whole cohort (OR=2.27) is increased (Fallerini C. et al., 2021 c).

In order to demonstrate the impact of polymorphism in increasing thrombotic risk, in a subset of 52 severely affected hospitalised males, four main laboratory parameters related to a pro-inflammatory state (lymphocyte count, D-dimer and LDH) and a higher risk for thrombosis (D-dimer, platelet count and LDH) were longitudinally followed.

The maximum pick (over 10 times of the normal upper value) was exclusive of Asp/Asn and Asn/ Asn genotypes and older patients. In addition, the pick timing was earlier in Asn/Asn (median 7.5 days from infection) than Asp/Asn (median 13.5 days from infection), (p value=3× 10–2), (**Figure 21, b-f**) (Fallerini C. et al., 2021 c). Since the vWF is a downstream effector for clotting, the non-0 blood groups, associating with more stable vWF, also correlate with higher D-dimer and LDH values (**Figure 21, g-h**), in line with previous reports (Severe Covid-19 GWAS Group et al., 2020).

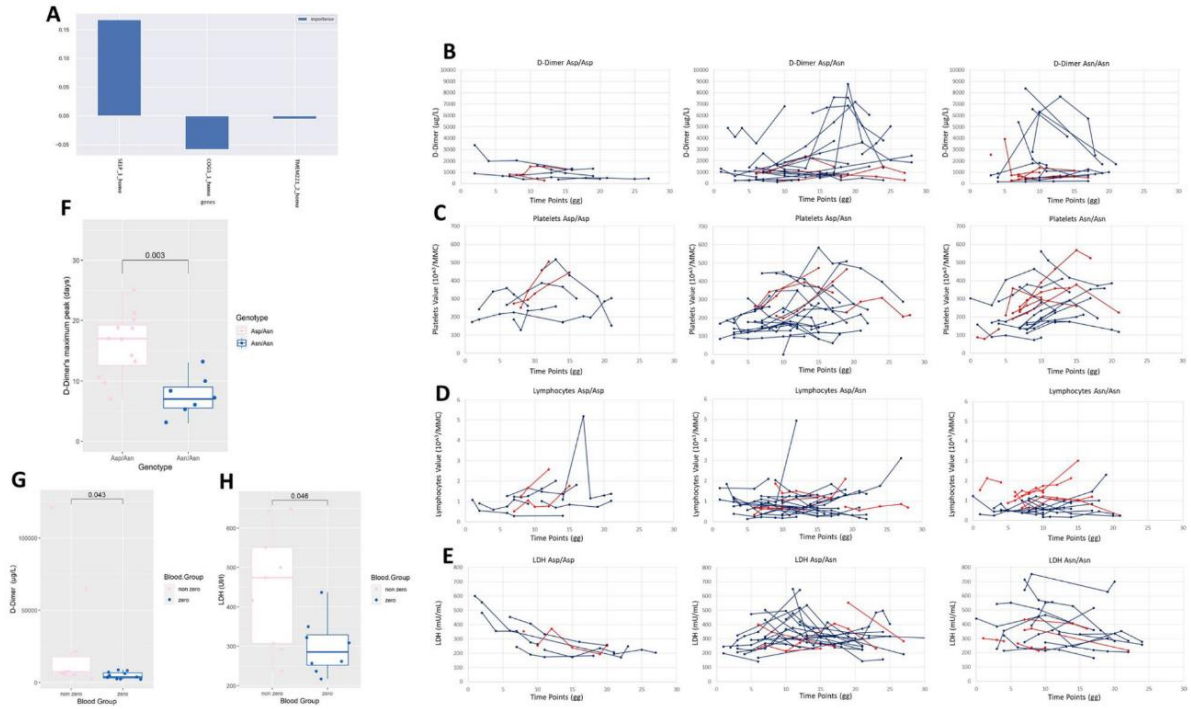


Figure 21. Homozygous genotype Asn/Asn at the polymorphic locus Asp603Asn (rs6127) is related to severity and to D-dimer pick. A Selection of *SELP* gene as relevant for severity. LASSO logistic regression on Boolean representation of homozygous common bi-allelic polymorphism of autosomal genes in males is presented. The LASSO logistic regression model provides an embedded feature selection method within the binary classification tasks (severe vs mild). The upward histogram means positive weights, i.e., the specific variant at the specific locus (feature) contributes to severity of COVID-19. SELP_1_homo=homozygous genotype Asn/ Asn at the polymorphic locus p.Asp603Asn (rs6127). The downward histograms mean negative weights, contributing to mildness of COVID-19. COG3_1_homo=homozygous genotype Ser/Ser at the polymorphic locus p.Leu825Ser (rs3014902). COG3 gene encodes for a vesicle docking protein involved in viral trafficking. TMEM221_2_homo=homozygous genotype Ala/Ala at the polymorphic locus p.Thr66Ala (rs4808641). TMEM221 gene encodes for a transmembrane protein. **B–E** Longitudinal laboratory data related to thrombosis and severity. Linear graphs of four laboratory values: D-dimer µg/L (B), platelets 103 /mmc (C), lymphocytes 103 /mmc (D), LDH UI/L (n.v. 135- 225 UI/L) (E). Each line represents each severe hospitalized patient. Each point represents the different time point (day) in which the different values have been measured. Patients aged ≥55 years are indicated in blue, while patients aged <55 are in red. **F** The D-dimer pick is earlier in the Asn/Asn (median=7.5 days) than the Asp/Asn genotype (p=3× 10–2 by Mann–Whitney test). Box plots of patients with D-dimer values above 2000 µg/l were represented. Only Asp/Asn (light blue) and Asn/Asn (pink) genotypes are represented because patients with the Asp/Asp genotype do not have the pick and do not show values above 2.000. **G, H** The nonzero group associates with higher D-dimer (G) and LDH values (H). Severe hospitalized patients with 0 blood group=light blue; non-0 blood group=pink in box plots. From Fallerini C. et al., 2021c.

Given the stronger association of the *SELP* polymorphism in older males, the *AR* poly-Q status would impact on the *SELP* genotype (Baldassarri M et al., 2021 b): the combination of poly-Q≥23 with homozygous *SELP* polymorphism versus D-dimer value reached an OR of 3.26 (Fallerini C. et a. 2021c). Therefore, the two polymorphisms enhance each other, being two pieces of the same puzzle contributing to thrombosis in COVID-19 males (**Figure 22**).

These results provide a rationale for the repurposing of antibodies against P-selectin, such as Inclacumab and Crizanlizumab, the latter as a prevention of vaso-occlusive crises in

patients with sickle cell disease, as add on therapy in rs6127 male homozygotes especially if older than 50 or with an impaired androgen receptor.

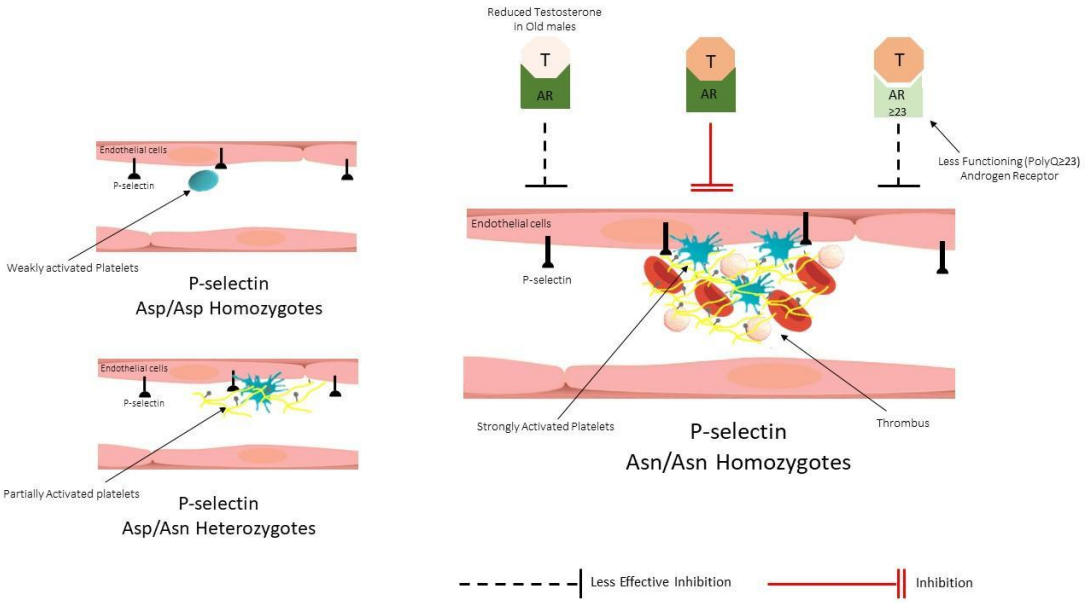


Figure 22. Combination of AR poly-Q ≥ 23 with homozygous *SELP* polymorphism. The homozygous Asn/Asn genotype for the Asp603Asn (rs6127) *SELP* polymorphism is correlated with increased platelet activation and risk of thrombosis. This effect is greater in elderly males or males with *AR* polyQ >23 in whom the effect of androgen inhibition on P-selectin is reduced (Fallerini, 2021a).

5.6 Post-Mendelian Model for predicting COVID-19 Severity.

The identification of host genetic factors modifying disease susceptibility and/or disease severity has the potential to reveal the biological basis of disease susceptibility and outcome as well as to subsequently contribute to treatment amelioration (Elhabyan et al. 2020).

COVID-19 represents a particularly interesting and accessible complex disorder for modelling host genetic data because the environmental factor (SARS-CoV-2) can be readily identified by a PCR-based swab test. The still moderate viral genome variability has thus far been shown to have relatively low impact on disease severity (Islam MR et al., 2020) where currently age, sex, and comorbidities are the major factors predicting disease susceptibility and outcome.

While these factors certainly have significant value for prediction, they provide limited insights into disease pathophysiology and are of limited relevance for drug development. Moreover, these variants only explain a small fraction of trait variability Moreover, these factors alone do not fully explain the clinical variability of the disease (Fallerini C. et al., 2022).

The Italian GEN-COVID Multicenter Study collected more than 5000 bio-specimens and clinical data from SARS-CoV-2-positive individuals (Daga S et al., 2021), and WES analysis contributed to the identification of rare variants (Fallerini et al. 2021a) and common polymorphisms (Baldassarri et al. 2021a; Croci et al. 2022; Fallerini et al. 2021b) associated with COVID-19 severity.

In order to explain COVID-19 as a complex disease, we therefore developed a model, called post-Mendelian, that would allow us to combine all the clinical variability identified by WES to predict the severity of each patient starting from the genetic data. This model allowed to combine all *common, low frequency, rare* and *ultra-rare* variants, extracted as implicated in COVID-19, in order to predict disease severity (**Figure 23**), (Fallerini C. et al 2022; Zguro K. et al., 2022).

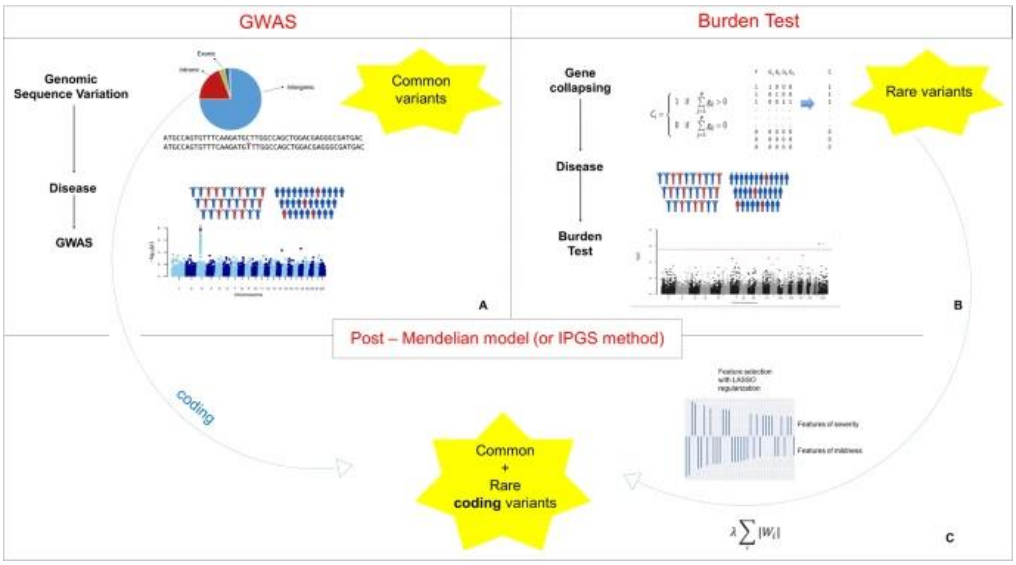


Figure 23. **Methodologies for features selection.** A Genome-wide association studies methodology for the study of SNPs. B Burden test methodology for the study of rare coding variants. C Post-Mendelian model for the study of both common and rare coding variants. From Zguro K. e al., 2022.

5.6.1 Definition of the Integrated PolyGenic Score (IPGS) to predict severity.

In order to reach a more precise disease severity prediction from genetic variability, within the post-Mendelian model, it was defined a score, called *Integrated PolyGenic Score* (IPGS), in which the Boolean representations were considered isolated from each other. The aim of the IPGS is to combine information from different representations to predict disease severity.

Assuming that variants with a low MAF, such as ultra-rare variants, have a stronger impact on the protein function in respect of variants with a high MAF, such as common

variants, the variables corresponding weights were defined based on their frequency. The Boolean features selected by the LASSO logistic models were used to calculate the variables in the formula below (**Figure 24**).

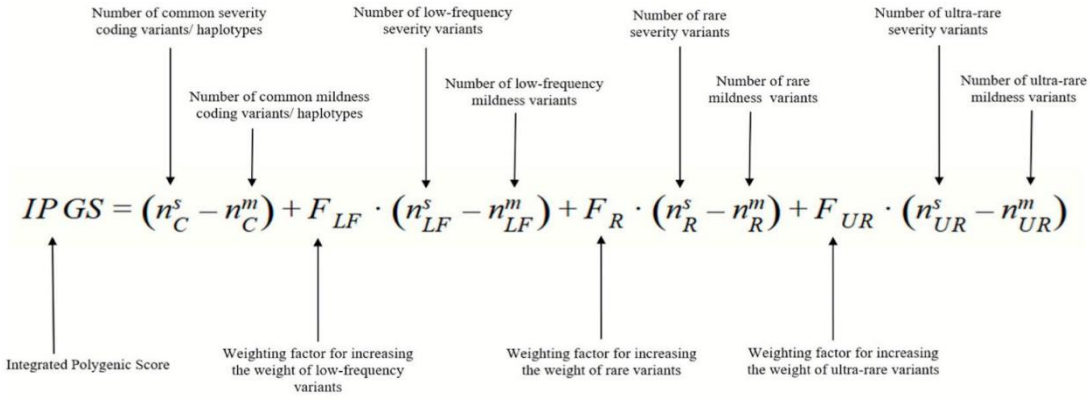


Figure 24. Graphic representation of the IPGS formula used for this model. From Fallerini C. et al., 2022.

In the formula (**Figure 24**), *n* variables are used to indicate the number of input features of the predictive model that promote the severe outcome (*superscript s*) or that protect from a severe outcome (*superscript m)* and with genetic variants having MAF ≥ 5% (common, *subscript C*), 1% < MAF ≤ 5% (low-frequency, *superscript LF*), 0.1% < MAF ≤ 1% (rare, *subscript R*), and MAF < 0.1% (ultra-rare, *subscript UR*). The weighting factors (F) $F_{LF, R}$, and F_{UR} were included to model the different penetrant effects of low-frequency, rare, and ultra-rare variants, compared to common variants. Thus, the 4 terms of the formula (Fig. 22) can be interpreted as the contributions of common, low-frequency, rare, and ultra-rare variants to a score that represents the genetic propensity of a patient to develop a severe form of COVID-19. The corresponding F variables were defined by optimizing the separation between severe and mild cases as offered by the IPGS formula. This optimization returned values of 2, 4, and 5 for the low-frequency, rare, and ultra-rare variants, respectively (Fallerini C. et al., 2022).

5.6.2 COVID-19 post-Mendelian model predictivity.

The procedure described in the previous sections completely defines how to calculate the IPGS (Figure 22). The predictive model of the binary COVID-19 severity (hospitalized patients with any form of respiratory support versus all other patients) was defined as a logistic model that uses as input features IPGS, age, and sex.

The predictive model has been defined as a logistic model that uses IPGS, age and gender as input characteristics. The method thus structured allowed to isolate the genetic factors underlying the severity of COVID-19, regardless of gender and age. Of course, only patients that deviates from their expected severity based on age and sex were used. The procedure was designed to isolate the genetic basis of COVID-19 severity.

Instead, in this final step, IPGS, age and sex are combined to predict the actual COVID-19 severity.

Moreover, to prevent overfitting, the model was fitted using 466 samples different from the training set. 3 logistic models were then created:

i) IPGS + age + sex;

ii) age + sex;

iii) only IPGS.

The increase in performances of the model with IPGS suggests that this score indeed confers significant additional (genetic) information for predicting COVID-19 severity compared to only age and sex (**Figure 25**), (Fallerini C. et al., 2022). The test set is composed of 466 patients not included in the training set previously exploited for the IPGS feature engineering. The model's performance was then tested using three independent cohort sets of European ancestry (UK, Germany, Canada) (**Figure 25A**). The model exhibited an overall accuracy of 0.73, precision equal to 0.78, with a sensitivity and specificity of 0.72 and 0.75, respectively. Of note, all the aforementioned metrics are higher than the corresponding values obtained using a logistic model that adopted as input features only age and sex.

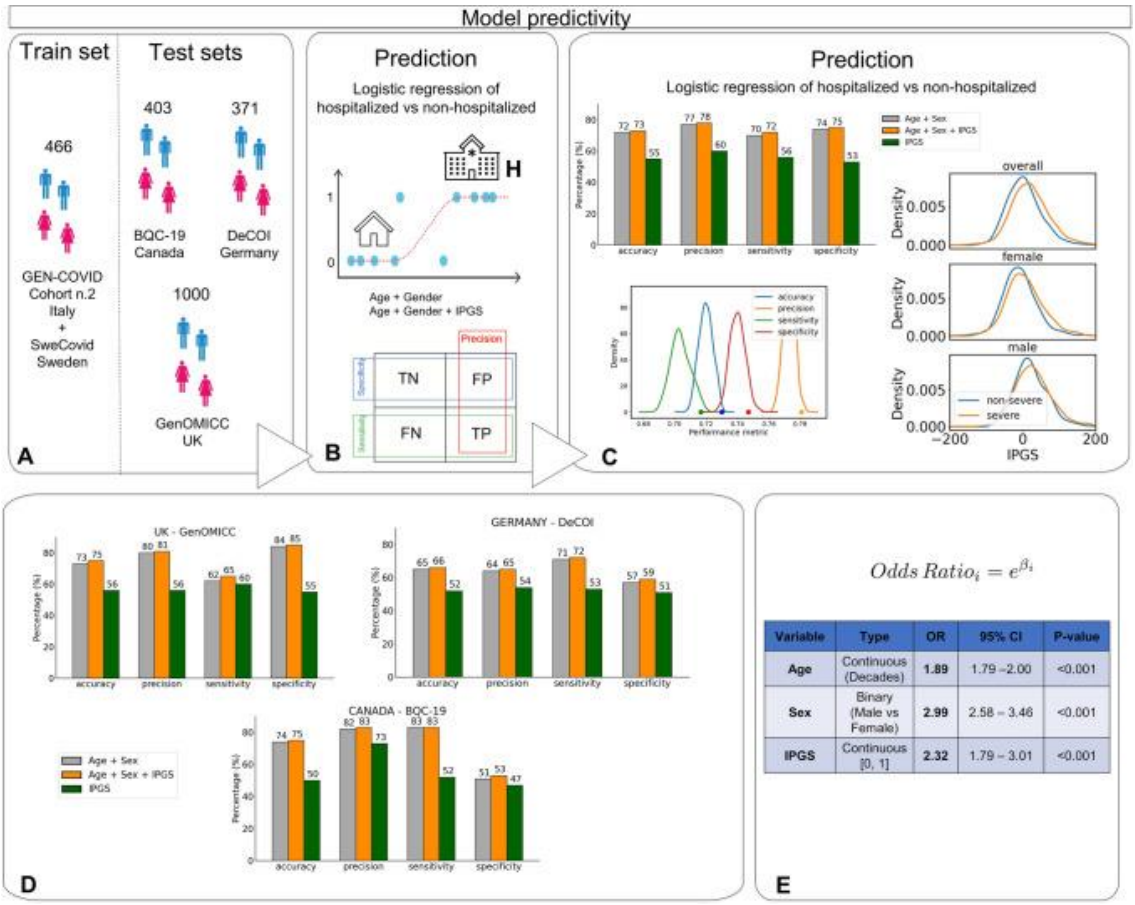
The increase in performances of the model with IPGS suggests that this score indeed confers significant additional (genetic) information for predicting COVID-19 severity compared to only age and sex. The increase of the performances results statistically significant (p value < 0.05 for accuracy, precision, sensitivity, specificity). Even the model fitted with IPGS alone shows performances well above the random guess. The increase in performance was systematically observed throughout all the cohorts: on average + 1.33% for accuracy, + 1% for precision, + 1.33% for sensitivity, + 1.67% for specificity.

Considering the difference in phenotype classification inherent to a comparison among various international cohorts, and the genetic variability among different European sub-populations, the consistent increase in performances observed for the model with IPGS demonstrates that this score provides a robust index for predicting COVID-19 severity.

As an additional test for the importance of the IPGS score to predict COVID-19 severity, the univariate logistic models were used on the overall set including both train and

test cohorts to estimate the OR of severe COVID-19 for IPGS, age, and sex, separately. The test confirmed that severity was associated with IPGS, showing an OR of 2.32 with age, measured in decades, and sex, having OR of 1.89 and 2.99 respectively.

When adjusting for comorbidities, in the train cohort where the comorbidities were available, with a multivariable logistic model, OR of IPGS was 2. This result further confirmed that IPGS is a reliable predictor of COVID-19 clinical severity (**Figure 25**), (Fallerini C. et al., 2022).



Variable	Type	OR	95% CI	P-value
Age	Continuous (Decades)	1.89	1.79 – 2.00	<0.001
Sex	Binary (Male vs Female)	2.99	2.58 – 3.46	<0.001
IPGS	Continuous [0, 1]	2.32	1.79 – 3.01	<0.001

Figure 25. Model predictivity. A) The post-Mendelian model was trained using a sample of 466 patients from the GEN-COVID cohort n.2 and Swedish cohort and tested with three additional European cohorts from UK, Germany and Canada. B) A logistic regression model was used for severity prediction. Severity was defined mainly on the basis of hospitalization versus not hospitalization. Hospitalized cases without respiratory support were included in controls. TN=true negative; TP= true positive; FN= false negative; FP= false positive. C) When the IPGS is added to age and gender, the performances of the model increase: accuracy +1%, precision +1%, sensitivity +2%, specificity +1%. These increases are statistically significant with respect to the null distribution obtained by randomizing the IPGS. The performances of the model built with IPGS alone are all above the random guess. In addition, on the right, we reported the distributions of the IPGS for severe and non-severe patients. D) In the three tested cohorts, when the IPGS is added to age and sex as a regressor, all the performances increase: the accuracy up to+2%, the precision up to+1%, the sensitivity up to+3%, and the specificity up to+2%. E) The univariate logistic regression models fitted on the cohort including both train and test, confirmed that the IPGS is associated with severity with an odds-ratio (OR) of 2.32, while age (continuous in decades) and sex have an OR of 1.89 and 2.99, respectively. From Fallerini C. et al., 2022.

5.6.3 Advantages of IPGS and clinical interpretability of connected features

In order to further demonstrate the predictive ability of IPGS, we compared the clinical outcome with the probability of severity obtained from the three different models: i) IPGS + age + sex; ii) age + sex; iii) only IPGS (**heatmap in Figure 26**). It appears evident that in a subset of patients, the 2 models based on sex-age alone and IPGS alone have a discordant prediction (**left and right end of dendrogram in Figure 26**). This is in accordance with the above-presented logistic regression analysis that shows IPGS having an OR of 2.32 for severity, identifying IPGS as a relevant predictor of severity.

Moreover, the features, on which the IPGS score is built, have also been proved to be important and significant from a biological point of view to identify pathophysiological mechanisms and possible personalized adjuvant treatments.

For example, three male patients, within two distinct age ranges (46–50, 51–55) (**panels B, C and D, Figure 26**) with severe disease (intubation and CPAP) are imperfectly represented by the sex-age model (probability of severity from 0.52 to 0.66) and better represented by the IPGS model (probability of severity from 0.91 to 0.95). The detected genetic variants that would allow to clinically consider putative personalized treatments in similar cases are: (1) *TLR7* ultra-rare variant indicating to consider possible adjuvant treatment with IFN-γ (Fallerini et al. 2021a); (2) homozygosity p.Asp603Asn in *SELP* gene suggesting putative adjuvant treatment with anti-selectin P autoantibodies such as Crizanlizumab (Fallerini et al. 2021b) and (3) polyQ longer than 23 in *AR* gene suggesting possible adjuvant treatment with testosterone (Baldassarri M et al., 2021a).

In a female patient, within age range 31–35, the sex-age model showed a probability of severity of 0.17 (**panel D, Figure 26**) while the IPGS score was 336 corresponding to a probability of severity of 0.95. The patient had no comorbidities except for hypothyroidism. She underwent steroid treatment and CPAP ventilation. She was found to be carrier of *ADAMTS13* ultra-rare variant. She had indeed a high D-dimer value, in line with a higher risk for thrombosis. Caplacizumab (an antibody anti-vWF) would be an option to consider as possible adjuvant treatment in the clinical management of similar cases.

Finally, two male patients, within two distinct age ranges (81–85, 86–90) (**panel F and G, Figure 26**) with a relatively mild respiratory disease (hospitalised with low-flow oxygen therapy) presented an IPGS score of 258 and 141, respectively. Their severity probabilities calculated on sex-age (0.9 and 0.94) do not mirror the relatively mild clinical outcome, which is instead better represented by the severity probability calculated in IPGS

only (0.23 and 0.41). Those two patients presented ultra-rare variants in *ACE2* gene, likely responsible for reduced viral load (Benetti et al., 2020a), and in *AGTR2* gene, which reduced activity is known to prevent cystic fibrosis pulmonary manifestation (Darrah et al., 2019).

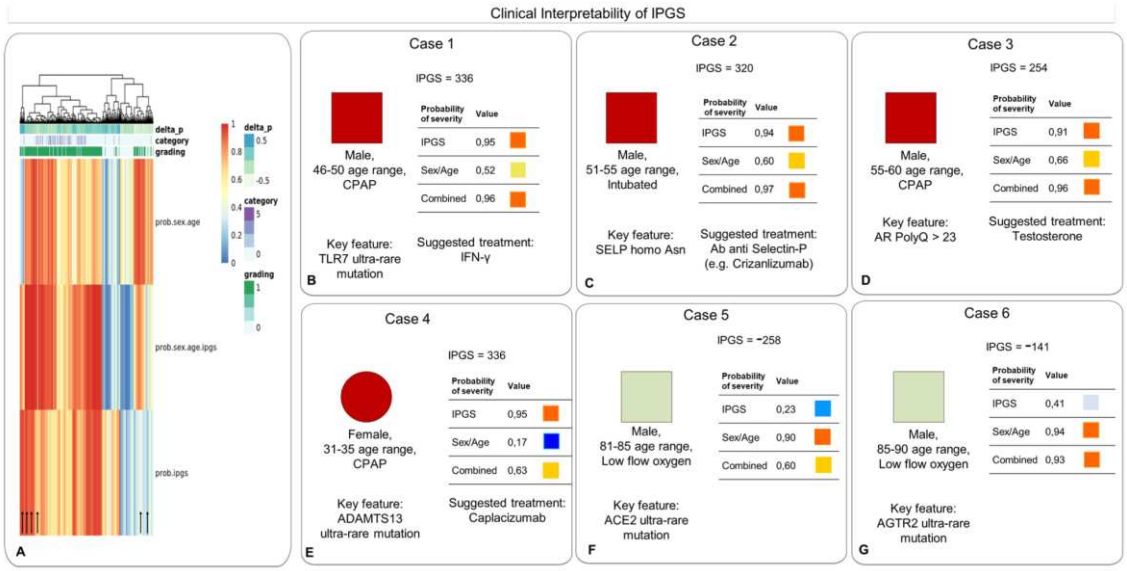


Figure 26. **Clinically interpretability of IPGS.** Panel A shows the GEN-COVID cohort dendrogram and heatmaps of the probabilities of severity based on the 3 different models: sex-age alone, IPGS alone and combined model. In the extreme ends of dendrogram (left and right) the probability of severity based on sex-age alone and IPGS alone is highly discordant (different colours). Selected examples corresponding to the arrows are illustrated in panels B-G. In each panel IPGS score, probabilities of severity and key features useful for bedside clinical management are shown. From Fallerini C. et al., 2022.

6. Discussion

COVID-19 pandemic represented a worldwide challenge in the past 3 years. Even if the emergency is ending, the identification of the main prognostic factors that determine the clinical outcome still remain one of the major challenges, to implement effective preventive and therapeutic strategies also useful for other infectious diseases and complex diseases. The disease, not only limited to the lungs, shall be defined as multi-organ with a high variability in symptoms; the majority of cases are mild (80%), with 15% progressing to severe pneumonia and 5% developing acute respiratory distress syndrome, sepsis, and/or multisystem organ failure.

The multisystem involvement is mainly due to the diffuse localization in the body of the *ACE2* receptors, the major viral entry point. It is known also that the severity of the disease is due to systemic inflammation and cytokine storm dysregulation. However, it is not yet clear which factors can cause the transition from a physiological response to a pathogenic hyper-inflammatory response and hence from mild to severe COVID-19. Age and comorbidities explain only part of this clinical variability. Thus, it was possible to hypothesize that host genetics may explain clinical differences of the disease. Therefore, it was mandatory to identify genetic variants linked to the severe disease, to the mild one, or even protectives.

To suit this need, since March 2020, a genetics-based approach to understand the clinical variability of COVID-19 has been created. To date, the Medical Genetics Unit of the University of Siena has collected biological samples of more than 5.000 Italian individuals infected by SARS-CoV-2 and related clinical data thanks to the foundation of the GEN-COVID consortium with actually more than 30 Italian hospitals, 16 local healthcare units and 8 departments of preventative medicine (Benetti E. et al., 2020 a; Benetti E. et al., 2020 b; Daga S et al., 2021).

Also thanks to International collaborations, it is now well known that also host genetics factors play a crucial role in causing the COVID-19 clinical outcome and various common and rare variants have been described in association with susceptibility and severity.

Concerning SARS-CoV-2 variants, although the majority of viral sequence's changes are projected to be detrimental and swiftly removed or to be neutral, a small number are predicted to influence functional characteristics, possibly modifying infection rate, disease severity, or interactions with the host immune system. Nonetheless, beginning in late

2020, the development of SARS-CoV-2 has been marked by the introduction of "variants of concern," or changes in viral properties such as disease transmission and antigenicity, most likely because of the changing immunological composition of the human species. In particular, large changes in the RBD region of the Omicron variant might contribute to high binding specificity with hACE2, which may result in a greater transmissibility, lower vaccine efficiency, and an increased risk of reinfection when compared to the Delta variant (Kumar et al., 2022). Otherwise, the viral genome variability has thus far been shown to have relatively low impact on disease clinical variability.

For this reason, the major factors involved in the severity of the disease remain the age, sex and host genetics.

COVID-19 has demonstrated itself to be a complex multifactorial disease, but its main environmental factor (SARS-CoV-2) is easily detectable by a PCR-base swab test. Thus, it represents an accessible disorder for identifying the role of human genetics in susceptibility to infection. These results have also further strengthened the hypothesis that COVID-19 is to be considered a complex and polygenic genetic disease and have helped to direct the development of an innovative model to represent this complexity.

However, there are two main gaps in the current knowledge concerning the impact of host genetics on the clinical manifestations of COVID-19. First, although several strong statistical associations have been identified between genes and disease severity, these have rarely been complemented by an interpretation of the biological mechanisms underlying the disease severity. In addition, host genetics predictive models of gravity, showed limited accuracy.

For this reason, the aim of the present study was to develop an easily interpretable model that could be used to predict the severity of COVID-19 from host genetic data. This approach would then provide useful information for the development of diagnostic and while being able to guide disease management at the bedside (Fallerini et al., 2022).

It had already been emerged the hypothesis that common variants could constitute a genetic background on which rare variants, with a greater impact on the protein, could act in determining the severity of the disease. (Benetti E. et al, 2020 b). Moreover, assuming that COVID-19 is a complex multisystem disease, and as such, a much greater number of genes are expected to be involved. Methods neglecting the combined contribution of common and rare variants were unlikely thus far to thoroughly characterize the host genetics underlying COVID-19. Thus, there were an urgency to create an organic model explaining how to combine all the genetic background to predict disease severity.

A novel predictive model, termed Post-Mendelian model, has been developed to aggregate the effects of all genetic components into a score, named Integrated PolyGenic Score (IPGS) (Fallerini et al., 2022). This procedure identified about 8000 characteristics, corresponding to about 4000 genes involved, confirming the hypothesis that COVID-19 was a complex disease. However, it also has been identified that a small fraction of severe cases was due to a Mendelian form.

It also emerged, immediately, that the differences observed in clinical manifestations between men and women were based on the genetic diversity between the two sexes, which was already known for other conditions. Epidemiological studies indicated that men and women had a similar susceptibility to SARS-CoV-2 infection, but the prognosis was worse in men. This model confirmed this hypothesis showing that 25% of the extracted genes were sex-specific. At the basis of this difference several pathogenic mechanisms were identified such as related to either genes located on the X chromosome or genes regulated in opposite directions by androgens and oestrogens when contributing with less penetrant common variants.

A strength of this model was to use oligo-asymptomatic SARS-CoV-2-infected subjects as controls instead of general population, providing a significant advantage in accuracy. Patients were indeed stratified based on COVID-19 severity using a modified version of the *WHO COVID-19 Outcome Scale* adjusted for sex and age.

It should be noted that, in most of the GWAS studies available in the literature, the general population (and not oligo-asymptomatic SARS-CoV-2-infected subjects) was used as a control. The use of the general population as a control was motivated by the need to increase the size of the dataset (ten/hundreds of thousands subjects), but this choice may have introduced artefacts in the analyses and thus affected the correct identification of genetic variants associated with the disease.

It has been possible to achieve the results of this study thanks to Italian GEN-COVID consortium that allowed to collect systematic detailed clinical and genetic data throughout all Italy, through a specific clinical questionnaire and the exome analysis centralized in the Genetic Unit of Siena Hospital. In addition, the involvement not only of several Italian hospitals but also of local healthcare units and departments of preventive medicine allowed to enrol also non-hospitalized oligo-asymptomatic, who were used, as already reported, as controls in this study.

The basis of the model is to predict a binary classification of patients into mild and severe cases, where a patient is considered severe if hospitalized and receiving any form of

respiratory support. At the same time, the selection of exome variants was another important step. The representation of the genetic variability by Boolean features responds to two requirements. Firstly, the usage of summary features at the gene level widely reduces the number of input features. Moreover, this combination of single genetic variants into gene-level variables highly facilitates the interpretation of the results. Interpretability is an important characteristic of ML models for predicting COVID-19 phenotypes, as only an easily interpretable model can be useful in clinical practice and significantly contribute to diagnostic and therapeutic targeting. The total number of input features with binary classification is much lower than the number of genetic variants, but still, Boolean features vastly outnumber the number of individual patients.

In this model, the conversion of genetic variants into Boolean features led to the definition of 12 separate sets of input features, based on their frequencies. It is important to keep in mind the observation that variants at different frequencies are expected to contribute differently to the phenotype, almost by definition. In this approach, common variants are indicated as common haplotypes since they are intended as combinations of coding variants within a single gene.

This post-Mendelian model has developed a new predictive score, called Integrated PolyGenic Score (IPGS), capable of predicting the severity of COVID-19 with an accuracy close to 73% (Fallerini C. et al., 2022). The implementation of the model and the verification of its performance was also possible thanks to international collaborations with working groups such as the WES/WGS Working Group within the HGI which allowed to test the approach on other non-Italian cohorts, further demonstrating its accuracy.

During these three years, thanks to this innovative approach, it has been possible the elucidation of numerous genetic markers that are not only likely to help in explaining the varied clinical outcomes of COVID-19 patients but can also guide the development of novel diagnostics and therapeutics.

As already state in the results section, the dual face of COVID-19 has been identified and demonstrated, which can manifest itself both as a *Mendelian form*, due to a single variant with a high impact on the protein, and as a *Polygenic form*, due to the combination of multiple variants with a low impact on the protein. In both forms, the identification of the genetic markers associated with the disease was fundamental to identify a possible add-on personalized therapy for the patient, through the repurposing of existing drugs or prioritization in vaccination campaigns.

The COVID-19 example has shown that the patient genetic characterization should become an essential and priority criterion to properly manage the patient also for other complex diseases.

Concerning the COVID-19 Mendelian form, three genetic markers has been identified, namely *TLR7*, *ADAMTS13* and *CFTR*, which were found to be responsible for severe COVID-19 in males, females and both sexes. Among these genes, it immediately emerges that the *TLR7* gene, located on the X chromosome, was identified only in males, while the *ADAMTS13* gene, subjected to regulation by oestrogen, was identified only in females. The association between *TLR7* ultra-rare variants and severe COVID-19 in males has been proved by the functional studies which have shown a reduced expression of TLR7-related genes in cases compared to controls and impairment in type I and II IFN responses. These findings elucidate the crucial role of TLR7 in the recognition of SARS-CoV-2 and in the following elicitation of an early antiviral immune response that could prevent the progress into a severe form of COVID-19 (Fallerini et al., 2021a; Mantovani S., et al. 2022). These data confirmed the *TLR7* role previously reported (van der Made C.I. et al. 2020) in severe disease in males, especially in young people (<60 years), where usually the severe symptoms are not so frequent.

Based on this result, a *TLR7* gene screening analysis could potentially be applied in male patients in order to define a better management and a personalized treatment. In particular, it would allow to consider an add-on treatment with IFN-γ administration (Mantovani S., et al. 2022, van Laarhoven A. et al., 2021).

The identification of the *ADAMTS13* gene as a genetic marker of severe disease in females allowed to highlight how, in addition to the immune system, other pathogenic mechanisms can also influence disease severity. Concerning *ADAMTS13*, in the presence of heterozygous variants, the underlying mechanism was the increased thrombotic risk. The penetrance of the thrombotic disease triggered by SARS-CoV-2 infection in heterozygous *ADAMTS13* subjects is incomplete. Furthermore, the disease segregates in families as an autosomal dominant disorder, conditioned by SARS-CoV-2 infection, sex, and age. Indeed, oestrogens really have the power to induce protein production (Powazniak Y. et al., 2011). For this reason, females from the puberal period until ovarian failure are protected by the action of oestrogens. In line with this, it has been identified that heterozygous females over 50 years of age are more at risk, as well as female paediatric cases, who also lose the beneficial effect of oestrogen. In the other sex (male), the period with less oestrogen is the one from puberty to andropause and therefore the tendency to micro-angiopathy is more

evident in males under 50 years of age. This finding has clinical relevance due to the availability of drugs such as Caplacizumab or Crizanlizumab that could be suggested to patients with *ADAMTS13* variants exhibiting decreased enzymatic activity. Caplacizumab, an anti-vWF bivalent single-domain nanobody, inhibits vWF–platelet interaction and is already used to treat acquired thrombotic thrombocytopenic purpura. Besides, Crizanlizumab is a monoclonal antibody that prevents leukocyte recruitment and platelet aggregation at the site of vascular damage by inhibiting P-selectin binding to its ligands. These two drugs are likely to replace the reduced activity of the metalloproteinase due to certain mutations and therefore they could also be useful in decreasing hyper-inflammation signs in heterozygous *ADAMTS13* patients (Zguro et al., 2022).

This study then highlighted the role of *CFTR* gene in COVID-19 disease severity. This gene has been widely studied and already known to be associated with an increased risk of recurrent infections in the case of heterozygous variants, as well as CF when bi-allelic variants are present. Among mechanically ventilated patients, CF-carriers were more represented (8.7%) and they were significantly ($p < 0.05$) younger (mean age 51 years) compared to non-carriers. Furthermore, in the whole cohort, the age of male CF-carriers was lower, compared to non-carriers ($p < 0.05$). CF-carriers had a relative risk of presenting an abnormal inflammatory response (CRP $\geq$ 20 mg/dL) of 1.69 ($p < 0.05$) and their hazard ratio of death at day 14 was 3.10 ($p < 0.05$) in a multivariate regression model, adjusted for age, sex and comorbidities. In conclusion, CF-carriers are more susceptible to the severe form of COVID-19, showing also higher risk of 14-day death (Baldassarri et al., 2021).

This finding provided a rationale for the repurposing of Ivacaftor (Kalydeco), the first CFTR modulator approved by the U.S. Food and Drug Administration (FDA), in patients carrying class III or IV *CFTR* gene variants.

The predisposition of CF-carriers to severe COVID-19 is probably related to multiple factors: the acidification of the airway surface liquid that impairs immune response (Shah V.S. et al, 2016); a reduced CFTR function that upregulates the proinflammatory signalling, and that is associated with a deficiency in pro-solving mediators, known to promote the resolution of the inflammation (Groman J.D. et al., 2004) and iii. an accumulation of misfolded CFTR that may trigger NF-kB signalling (Sarantis P. et al., 2020). In addition, as further confirmation of this result, it was highlighted that the geographical distribution of CF carriers is related to the spread and mortality of COVID-19 in 37 countries (Gabbi C. et al, 2022).

Whereas, it's interesting to notice that carriers of bi-allelic CF-causing variants (CF patients), undergo, instead, a mild form of COVID-19 (Bezzerri V. et al., 2020; Colombo C. et al., 2020; Cosgriff R. et al., 2020). The explanation may lay the fact that CF patients, while costumed to always wear protective masks, are often treated with modulators that re-establish CFTR function or with other drugs, like azithromycin, that may protect against infections (Bezzerri V. et al., 2020; Colombo C. et al., 2020).

Moreover, the machine learning post-Mendelian model pinpointed *CFTR* as a bidirectional modulator of COVID-19 outcomes. In particular, the rare complex allele [G576V;R668C] is associated with a milder disease via a gain-of-function mechanism. Conversely, *CFTR* ultra-rare alleles with reduced function are associated with disease severity either alone (dominant disorder) or with another hypomorphic allele in the second chromosome (recessive disorder) with a global residual CFTR activity between 50 to 91%. In conclusion, *CFTR* genetic analysis is an important tool in identifying patients at risk of severe COVID-19 the global impairment of CFTR function, either by a strong LOF allele in heterozygosity or compound heterozygosity for two hypomorphic alleles, is associated with COVID-19 severity (Baldassarri et al., 2022).

Since Mendelian forms of COVID-19 are able to explain only a small fraction of severe cases, it has been investigated the role of some common polymorphisms in determining the severe manifestation of the disease. Among these, the first identified was the polymorphism polyQ tract of the *AR* gene, in males. *AR* polyQ length correlates with receptor functionality, with shorter polymorphic glutamine repeats typically associated with higher and longer polyQ tracts with lower receptor activity. AR is expressed in both males and females, but the bioavailability of its ligands T and dihydroT (DHT) differs significantly, being much higher in males. As previous studies linked male hypogonadism to a poorer outcome in COVID-19 patients, it has been demonstrated that shorter polymorphic glutamine repeats (≤22) confer protection against life-threatening COVID-19 in a subpopulation of individuals with age <60 years. Specifically, longer polyQ size (≥23) is associated with higher serum T levels, suggestive of impaired negative feedback at the level of the hypothalamus and pituitary gland (Baldassarri et al., 2021 b). An improvement in peak oxygen saturation in men receiving T replacement therapy has been demonstrated in a randomized controlled trial and could be one of the mechanisms responsible for the observed protective effect of AR's with shorter polyQ tract in COVID-19 patients. For this reason, a simple genetic test measuring the *AR* polyQ repeat can be used in male patients to screen for those who are more likely to benefit from T therapy (Baldassarri et al., 2021 b).

Moreover, concerning the cascade of sexual hormones, *CYP19A1* gene, encoding for the testosterone-to-estradiol metabolizing enzyme CYP19A1 (also known as aromatase), has been identified as a host factor that contributes to worsened disease outcome in SARS-CoV-2-infected males. The causal role of *CYP19A1* in male SARS-CoV-2 pathogenesis and male lung health was underpinned by treating infected hamsters with the clinically approved CYP19A1 inhibitor letrozole. Male and female hamsters treated with letrozole showed elevated circulating testosterone levels and reduced circulating estradiol levels suggestive of functional CYP19A1 inhibition in both sexes. Consequently, letrozole-treated males showed improved lung function that was reflected by ameliorated lung pathology and accelerated weight gain after virus clearance. In contrast, letrozole treatment did not show any benefit for SARS-CoV-2-infected females (Stanelle-Bertram et al., 2023).

So, while *AR*-related findings might suggest that replacement of testosterone levels could provide an add-on treatment option, this should be considered with caution because of the complexity of testosterone treatment in infected animals, which could be beneficial or detrimental depending on the dosage used. However, the detection of testosterone-dependent CYP19A1 expression further highlights the promising role of CYP19A1-targeted therapy that aims to balance sex hormone levels rather than substituting one hormone alone. Thus, patients with additional sex hormone balance-disturbing conditions (longer *AR* polyQ size, CYP*19A1* Thr201Met, age, and obesity) might additionally benefit from letrozole treatment upon SARS-CoV-2 infection (Stanelle-Bertram et al., 2023).

Another polymorphism identified in males was p.Leu412phe in the *TLR3* gene as a marker of severity. This polymorphism makes males, in whom after puberty testosterone lowers TLR3 expression, at risk of severe COVID-19 in a context of a polygenic model. Moreover, based on impairment of autophagy, these data provide a rationale for reinterpreting clinical trials with HCQ stratifying patients by L412F. In particular, it has been defined an important role of autophagy downstream of the TLR3 receptor, possibly affecting TNF production and susceptibility to infections, including SARS-CoV-2, pinpointing to IFNs treatment (especially IFN-γ) avoiding HCQ.

Among *SELP* variants, the p.Asp603Asn functional polymorphism has been associated with thrombotic risk in various conditions. The polymorphism has indeed been shown to affect the binding of P-selectin to its ligand on leukocytes, possibly making the protein more efficient at recruiting leukocytes to the endothelium. The hyper-inflammatory and hyper-thrombotic state, due to viral injury of the vascular endothelium, leads to the release of P-selectin by activated platelets, driving thrombosis and vascular inflammation

probably more efficiently in those individuals with enhanced P-selectin activities due the specific polymorphisms in homozygous (Tregouet D.A., 2002). Since *SELP* transcription is inhibited by androgens, the strength of the association should increase with age. Given the stronger association of the *SELP* polymorphism in older males, the *AR* poly-Q status would impact on the *SELP* genotype: the combination of poly-Q≥23 with homozygous *SELP* polymorphism versus D-dimer value reached an OR of 3.26. This result indicates that the two polymorphisms enhance each other, being two pieces of the same puzzle contributing to thrombosis in COVID-19 males (Fallerini et al., 2021 b). These results provide a rationale for the repurposing of antibodies against P-selectin, Crizanlizumab, as add-on therapy in p.Asp603Asn male homozygotes especially if older than 50 or with an impaired androgen receptor (Fallerini, 2021 b).

Moreover, concerning Long-COVID, two genes have been, for the moment, identified. For instance, it has been demonstrated that the combination of L412F in TLR3 and specific HLA class II haplotypes may put male patients at risk of post-acute sequelae of SARS-CoV-2 infection pointing to the need for an appropriate follow-up (Croci S. et al., 2022). Furthermore, it has been shown that ultra-rare variants of *RTEL1*, besides a predictive marker of COVID-19 severity, could be considered as a marker of pathological evolution of pulmonary fibrosis in the post-COVID phase (Bergantini L. et al., 2023). These notions should be used for rapid screening in hospitalized infected people and for appropriate follow-up assessment in at-risk individuals.

These results together demonstrate how this new approach, called the post-Mendelian model, has identified common, low-frequency, rare and ultra-rare variants associated with disease severity. Modelling precisely the role of the entire range of host genomics affecting disease susceptibility and severity in COVID-19 is critical to obtaining a complete biological understanding of the aetiology and pathogenicity of COVID-19 as well as other severe complex diseases. IPGS relies on both polymorphisms and rare variants is capable of differentially weighting features in an indirectly proportional way in respect to frequency, and therefore, to protein impact. Each patient indeed is assigned both a number and the list of her/his common and low-frequency polymorphisms relevant to COVID-19 supported by medically actionable information and of rare and ultra-rare variants conferring either risk of severity or protection from severe disease. Drawing on the entire picture presented through IPGS analysis, personalized adjuvant therapy could be envisaged. (Fallerini C. et al. 2022).

7. Conclusions and future perspectives

A better understanding of the role of host genetics and the application of a model that allows for the early identification of those individuals at risk of severe forms of the disease, will allow to activate specific measures that can reduce the impact of the disease on the community and to set up a dedicated patient care of the patient and a personalized therapy, through the "*repurposing"* of existing drugs or "*prioritization*" in vaccination campaigns. The identification of patient's genetic "fingerprint" should therefore become an essential and priority criterion to correctly take care of the patient himself, in COVID-19 as well as in other possible infectious or in other complex diseases.

For this reason, stratifying patients based on the genetic markers described in this study, both ultra-rare variants and common polymorphisms, it could be speculated the use of the drugs proposed as an add-on therapy.

The functional interpretation of the variants identified by the feature selection approach, complemented by the strong link between the involved human biological pathways and COVID-19 pathogenicity, supports the hypothesis that the IPGS score may contribute significantly to predicting the severity of COVID-19.

Indeed, there was a significant overall increase in performance compared to the model based solely on age and sex. This concept indicates that IPGS is a novel prognostic factor that should be considered in the management of COVID-19 patients.

The application of IPGS based on Machine Learning principles within a post-Mendelian model allows to more precisely identify the gene variants involved in COVID-19 as well as their specific roles, individually and in combination. The in-depth knowledge of genetics that allows, contributes or even helps to prevent diseases is fundamental and directly translatable into the development of personalized drugs as well as specific prevention and treatment protocols.

Finally, studies are currently ongoing aimed at elucidating deeper the genetic basis of susceptibility to COVID-19 severe forms in the paediatric population and to the development of Long-COVID. Similarly, possible genetic factors underlying infection resistance, variability in vaccination response and multi-organ phenotype of the disease are under investigation.

Bibliografia

Andolfo, Immacolata et al. “Common variants at 21q22.3 locus influence *MX1* and *TMPRSS2* gene expression and susceptibility to severe COVID-19.” *iScience* vol. 24,4 (2021): 102322. doi:10.1016/j.isci.2021.102322

Ashraf, Usman M et al. “SARS-CoV-2, ACE2 expression, and systemic organ invasion.” *Physiological genomics* vol. 53,2 (2021): 51-60. doi:10.1152/physiolgenomics.00087.2020

Ather, Amber et al. “Coronavirus Disease 19 (COVID-19): Implications for Clinical Dental Care.” *Journal of endodontics* vol. 46,5 (2020): 584-595. doi:10.1016/j.joen.2020.03.008

Ay, Cihan et al. “P-selectin gene haplotypes modulate soluble P-selectin concentrations and contribute to the risk of venous thromboembolism.” *Thrombosis and haemostasis* vol. 99,5 (2008): 899-904. doi:10.1160/TH07-11-0672

Baldassarri, Margherita et al. “Severe COVID-19 in Hospitalized Carriers of Single *CFTR* Pathogenic Variants.” *Journal of personalized medicine* vol. 11,6 558. 15 Jun. 2021, doi:10.3390/jpm11060558

Baldassarri, Margherita et al. “Gain- and Loss-of-Function *CFTR* Alleles Are Associated with COVID-19 Clinical Outcomes.” *Cells* vol. 11,24 4096. 16 Dec. 2022, doi:10.3390/cells11244096

Baldassarri, Margherita et al. “Shorter androgen receptor polyQ alleles protect against life-threatening COVID-19 disease in European males.” *EBioMedicine* vol. 65 (2021): 103246. doi:10.1016/j.ebiom.2021.103246

Blann, Andrew D et al. “The adhesion molecule P-selectin and cardiovascular disease.” *European heart journal* vol. 24,24 (2003): 2166-79. doi:10.1016/j.ehj.2003.08.021

Benetti, Elisa et al. “ACE2 gene variants may underlie interindividual variability and susceptibility to COVID-19 in the Italian population.” *European journal of human genetics : EJHG* vol. 28,11 (2020): 1602-1614. doi:10.1038/s41431-020-0691-z

Benetti, Elisa et al. “Clinical and molecular characterization of COVID-19 hospitalized patients.” *PloS one* vol. 15,11 e0242534. 18 Nov. 2020, doi:10.1371/journal.pone.0242534

Bergantini, Laura et al. “Ultra-rare RTEL1 gene variants associate with acute severity of COVID-19 and evolution to pulmonary fibrosis as a specific long COVID disorder.” *Respiratory research* vol. 24,1 158. 16 Jun. 2023, doi:10.1186/s12931-023-02458-7

Bezzerri, Valentino et al. “Does cystic fibrosis constitute an advantage in COVID-19 infection?.” *Italian journal of pediatrics* vol. 46,1 143. 6 Oct. 2020, doi:10.1186/s13052-020-00909-1

Bohn, Mary Kathryn et al. “Pathophysiology of COVID-19: Mechanisms Underlying Disease Severity and Progression.” *Physiology (Bethesda, Md.)* vol. 35,5 (2020): 288-301. doi:10.1152/physiol.00019.2020

Boyle, Evan A et al. “An Expanded View of Complex Traits: From Polygenic to Omnigenic.” *Cell* vol. 169,7 (2017): 1177-1186. doi:10.1016/j.cell.2017.05.038

Bongiovanni, Dario et al. “SARS-CoV-2 infection is associated with a pro-thrombotic platelet phenotype.” *Cell death & disease* vol. 12,1 50. 5 Jan. 2021, doi:10.1038/s41419-020-03333-9

Callewaert, Leen et al. “Implications of a polyglutamine tract in the function of the human androgen receptor.” *Biochemical and biophysical research communications* vol. 306,1 (2003): 46-52. doi:10.1016/s0006-291x(03)00902-1

Caracciolo, Massimo et al. “Efficacy and Effect of Inhaled Adenosine Treatment in Hospitalized COVID-19 Patients.” *Frontiers in immunology* vol. 12 613070. 18 Mar. 2021, doi:10.3389/fimmu.2021.613070

Carvalho-Schneider, Claudia et al. “Follow-up of adults with noncritical COVID-19 two months after symptom onset.” *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases* vol. 27,2 (2021): 258-263. doi:10.1016/j.cmi.2020.09.052

Carlos, W Graham et al. “Novel Wuhan (2019-nCoV) Coronavirus.” *American journal of respiratory and critical care medicine* vol. 201,4 (2020): P7-P8. doi:10.1164/rccm.2014P7

Casanova, Jean-Laurent et al. “Human TLRs and IL-1Rs in host defense: natural insights from evolutionary, epidemiological, and clinical genetics.” *Annual review of immunology* vol. 29 (2011): 447-91. doi:10.1146/annurev-immunol-030409-101335

Cervantes-Barragan, Luisa et al. “Control of coronavirus infection through plasmacytoid dendritic-cell-derived type I interferon.” *Blood* vol. 109,3 (2007): 1131-7. doi:10.1182/blood-2006-05-023770

Channappanavar, Rudragouda et al. “IFN-I response timing relative to virus replication determines MERS coronavirus infection outcomes.” *The Journal of clinical investigation* vol. 129,9 3625-3639. 29 Jul. 2019, doi:10.1172/JCI126363

Chung, Jaeyoon et al. “Comparison of methods for multivariate gene-based association tests for complex diseases using common variants.” *European journal of human genetics : EJHG* vol. 27,5 (2019): 811-823. doi:10.1038/s41431-018-0327-8

Colombo, Carla et al. “Impact of COVID-19 on people with cystic fibrosis.” *The Lancet. Respiratory medicine* vol. 8,5 (2020): e35-e36. doi:10.1016/S2213-2600(20)30177-6

COVID-19 Host Genetics Initiative. “Mapping the human genetic architecture of COVID-19.” *Nature* vol. 600,7889 (2021): 472-477. doi:10.1038/s41586-021-03767-x

COVID-19 Host Genetics Initiative. “A first update on mapping the human genetic architecture of COVID-19.” *Nature* vol. 608,7921 (2022): E1-E10. doi:10.1038/s41586-022-04826-7

Croci, Susanna et al. “The polymorphism L412F in *TLR3* inhibits autophagy and is a marker of severe COVID-19 in males.” *Autophagy* vol. 18,7 (2022): 1662-1672. doi:10.1080/15548627.2021.1995152

Cui, Jie et al. “Origin and evolution of pathogenic coronaviruses.” *Nature reviews. Microbiology* vol. 17,3 (2019): 181-192. doi:10.1038/s41579-018-0118-9

Daga, Sergio et al. “Employing a systematic approach to biobanking and analyzing clinical and genetic data for advancing COVID-19 research.” *European journal of human genetics : EJHG* vol. 29,5 (2021): 745-759. doi:10.1038/s41431-020-00793-7

Danecek, Petr et al. “Twelve years of SAMtools and BCFtools.” *GigaScience* vol. 10,2 (2021): giab008. doi:10.1093/gigascience/giab008

Darrah, Rebecca J et al. “AGTR2 absence or antagonism prevents cystic fibrosis pulmonary manifestations.” *Journal of cystic fibrosis : official journal of the European Cystic Fibrosis Society* vol. 18,1 (2019): 127-134. doi:10.1016/j.jcf.2018.05.013

Drosten, Christian et al. “Transmission of MERS-coronavirus in household contacts.” *The New England journal of medicine* vol. 371,9 (2014): 828-35. doi:10.1056/NEJMoa1405858

Eendebak, Robert J A H et al. “The androgen receptor gene CAG repeat in relation to 4-year changes in androgen-sensitive endpoints in community-dwelling older European men.” *European journal of endocrinology* vol. 175,6 (2016): 583-593. doi:10.1530/EJE-16-0447

Elhabyan, Abdelazeem et al. “The role of host genetics in susceptibility to severe viral infections in humans and insights into host genetics of severe COVID-19: A systematic review.” *Virus research* vol. 289 (2020): 198163. doi:10.1016/j.virusres.2020.198163

Elborn, J Stuart. “Cystic fibrosis.” *Lancet (London, England)* vol. 388,10059 (2016): 2519-2531. doi:10.1016/S0140-6736(16)00576-6

Fallerini, Chiara et al. “Association of Toll-like receptor 7 variants with life-threatening COVID-19 disease in males: findings from a nested case-control study.” *eLife* vol. 10 e67569. 2 Mar. 2021, doi:10.7554/eLife.67569

Fallerini, Chiara et al. “Common, low-frequency, rare, and ultra-rare coding variants contribute to COVID-19 severity.” *Human genetics* vol. 141,1 (2022): 147-173. doi:10.1007/s00439-021-02397-7

Fallerini, Chiara et al. “SELP Asp603Asn and severe thrombosis in COVID-19 males.” *Journal of hematology & oncology* vol. 14,1 123. 16 Aug. 2021, doi:10.1186/s13045-021-01136-9

Forni, Diego et al. “Molecular Evolution of Human Coronavirus Genomes.” *Trends in microbiology* vol. 25,1 (2017): 35-48. doi:10.1016/j.tim.2016.09.001

Franco, Luis H et al. “Autophagy downstream of endosomal Toll-like receptor signaling in macrophages is a key mechanism for resistance to *Leishmania major* infection.” *The Journal of biological chemistry* vol. 292,32 (2017): 13087-13096. doi:10.1074/jbc.M117.780981

Gabbi, Chiara et al. “Geographical distribution of cystic fibrosis carriers as population genetic determinant of COVID-19 spread and fatality in 37 countries.” *The Journal of infection* vol. 85,3 (2022): 318-321. doi:10.1016/j.jinf.2022.06.006

Graham, Rachel L et al. “A decade after SARS: strategies for controlling emerging coronaviruses.” *Nature reviews. Microbiology* vol. 11,12 (2013): 836-48. doi:10.1038/nrmicro3143

Groman, Joshua D et al. “Variation in a repeat sequence determines whether a common variant of the cystic fibrosis transmembrane conductance regulator gene is pathogenic or benign.” *American journal of human genetics* vol. 74,1 (2004): 176-9. doi:10.1086/381001

Gubbels Bupp, Melanie R, and Trine N Jorgensen. “Androgen-Induced Immunosuppression.” *Frontiers in immunology* vol. 9 794. 17 Apr. 2018, doi:10.3389/fimmu.2018.00794

Guo, Michael H et al. “Burden Testing of Rare Variants Identified through Exome Sequencing via Publicly Available Control Data.” *American journal of human genetics* vol. 103,4 (2018): 522-534. doi:10.1016/j.ajhg.2018.08.016

Hafez, Wael et al. “Reduced ADAMTS13 Activity in Correlation with Pathophysiology, Severity, and Outcome of COVID-19: A Retrospective Observational Study.” *International journal of infectious diseases : IJID : official publication of the International Society for Infectious Diseases* vol. 117 (2022): 334-344. doi:10.1016/j.ijid.2022.02.019

Hägg, Sara et al. “Gene-based meta-analysis of genome-wide association studies implicates new loci involved in obesity.” *Human molecular genetics* vol. 24,23 (2015): 6849-60. doi:10.1093/hmg/ddv379

Haybar, Habib et al. “Mutations and Common Polymorphisms in ADAMTS13 and vWF Genes Responsible for Increasing Risk of Thrombosis.” *Cardiovascular & hematological disorders drug targets* vol. 18,3 (2018): 176-181. doi:10.2174/1871529X18666180419102214

Helms, Julie et al. “High risk of thrombosis in patients with severe SARS-CoV-2 infection: a multicenter prospective cohort study.” *Intensive care medicine* vol. 46,6 (2020): 1089-1098. doi:10.1007/s00134-020-06062-x

Hoebe, Kasper et al. “The interface between innate and adaptive immunity.” *Nature immunology* vol. 5,10 (2004): 971-4. doi:10.1038/ni1004-971

Hoffmann, Markus et al. “SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor.” *Cell* vol. 181,2 (2020): 271-280.e8. doi:10.1016/j.cell.2020.02.052

Hui, David S et al. “Middle East respiratory syndrome coronavirus: risk factors and determinants of primary, household, and nosocomial transmission.” *The Lancet. Infectious diseases* vol. 18,8 (2018): e217-e227. doi:10.1016/S1473-3099(18)30127-0

Islam, M Rafiul et al. “Genome-wide analysis of SARS-CoV-2 virus strains circulating worldwide implicates heterogeneity.” *Scientific reports* vol. 10,1 14004. 19 Aug. 2020, doi:10.1038/s41598-020-70812-6

Janeway, Charles A Jr, and Ruslan Medzhitov. “Innate immune recognition.” *Annual review of immunology* vol. 20 (2002): 197-216. doi:10.1146/annurev.immunol.20.083001.084359

Keller, Christian W, and Jan D Lünemann. “Autophagy and Autophagy-Related Proteins in CNS Autoimmunity.” *Frontiers in immunology* vol. 8 165. 27 Feb. 2017, doi:10.3389/fimmu.2017.00165

Kosmicki, Jack A et al. “Pan-ancestry exome-wide association analyses of COVID-19 outcomes in 586,157 individuals.” *American journal of human genetics* vol. 108,7 (2021): 1350-1355. doi:10.1016/j.ajhg.2021.05.017

Kousathanas, Athanasios et al. “Whole-genome sequencing reveals host factors underlying critical COVID-19.” *Nature* vol. 607,7917 (2022): 97-103. doi:10.1038/s41586-022-04576-6

Kumar, Suresh et al. “Omicron and Delta variant of SARS-CoV-2: A comparative computational study of spike protein.” *Journal of medical virology* vol. 94,4 (2022): 1641-1649. doi:10.1002/jmv.27526

Lai, Jiann-Jyh et al. “Androgen receptor influences on body defense system via modulation of innate and adaptive immune systems: lessons from conditional AR knockout mice.” *The American journal of pathology* vol. 181,5 (2012): 1504-12. doi:10.1016/j.ajpath.2012.07.008

Levine, Beth et al. “Autophagy in immunity and inflammation.” *Nature* vol. 469,7330 (2011): 323-35. doi:10.1038/nature09782

Liaw A, Wiener M. Classification and regression by randomForest. R News. 2002. 2(3):18–22.

Li, Heng, and Richard Durbin. “Fast and accurate long-read alignment with Burrows-Wheeler transform.” *Bioinformatics (Oxford, England)* vol. 26,5 (2010): 589-95. doi:10.1093/bioinformatics/btp698

Loo, Joan et al. “COVID-19, immunothrombosis and venous thromboembolism: biological mechanisms.” *Thorax* vol. 76,4 (2021): 412-420. doi:10.1136/thoraxjnl-2020-216243

Luk, Hayes K H et al. “Molecular epidemiology, evolution and phylogeny of SARS coronavirus.” *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases* vol. 71 (2019): 21-30. doi:10.1016/j.meegid.2019.03.001

Mahmudpour, Mehdi et al. “COVID-19 cytokine storm: The anger of inflammation.” *Cytokine* vol. 133 (2020): 155151. doi:10.1016/j.cyto.2020.155151

Manne, Bhanu Kanth et al. “Platelet gene expression and function in patients with COVID-19.” *Blood* vol. 136,11 (2020): 1317-1329. doi:10.1182/blood.2020007214

Mantovani, Stefania et al. “Rare variants in Toll-like receptor 7 results in functional impairment and downregulation of cytokine-mediated signaling in COVID-19 patients.” *Genes and immunity* vol. 23,1 (2022): 51-56. doi:10.1038/s41435-021-00157-1

Marouli, Eirini et al. “Rare and low-frequency coding variants alter human adult height.” *Nature* vol. 542,7640 (2017): 186-190. doi:10.1038/nature21039

Mathew, Hannah R et al. “Systematic review: cystic fibrosis in the SARS-CoV-2/COVID-19 pandemic.” *BMC pulmonary medicine* vol. 21,1 173. 20 May. 2021, doi:10.1186/s12890-021-01528-0

Matsumoto, Misako et al. “Antiviral responses induced by the TLR3 pathway.” *Reviews in medical virology* vol. 21,2 (2011): 67-77. doi:10.1002/rmv.680

McCoy, J et al. “Androgen receptor genetic variant predicts COVID-19 disease severity: a prospective longitudinal study of hospitalized COVID-19 male patients.” *Journal of the European Academy of Dermatology and Venereology : JEADV* vol. 35,1 (2021): e15-e17. doi:10.1111/jdv.16956

Medzhitov, R, and C Janeway Jr. “Innate immunity.” *The New England journal of medicine* vol. 343,5 (2000): 338-44. doi:10.1056/NEJM200008033430506

Merten, M, and P Thiagarajan. “P-selectin in arterial thrombosis.” *Zeitschrift fur Kardiologie* vol. 93,11 (2004): 855-63. doi:10.1007/s00392-004-0146-5

Miao, Yi et al. “Potential Treatments for COVID-19 Related Cytokine Storm - Beyond Corticosteroids.” *Frontiers in immunology* vol. 11 1445. 16 Jun. 2020, doi:10.3389/fimmu.2020.01445

Millar, Adam C et al. “Predicting low testosterone in aging men: a systematic review.” *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne* vol. 188,13 (2016): E321-E330. doi:10.1503/cmaj.150262

Miller, Aaron C et al. “Cystic fibrosis carriers are at increased risk for a wide range of cystic fibrosis-related conditions.” *Proceedings of the National Academy of Sciences of the United States of America* vol. 117,3 (2020): 1621-1627. doi:10.1073/pnas.1914912117

Mohamad, Nur-Vaizura et al. “The relationship between circulating testosterone and inflammatory cytokines in men.” *The aging male : the official journal of the International Society for the Study of the Aging Male* vol. 22,2 (2019): 129-140. doi:10.1080/13685538.2018.1482487

Mohan, Shalini S et al. “Higher serum testosterone and dihydrotestosterone, but not oestradiol, are independently associated with favourable indices of lung function in community-dwelling men.” *Clinical endocrinology* vol. 83,2 (2015): 268-76. doi:10.1111/cen.12738

Molla M, et al. Using machine learning to design and interpret gene-expression microarrays. AIMag. 2004;25:23

Mukherjee, Suprabhat et al. “Toll-like receptor polymorphism in host immune response to infectious diseases: A review.” *Scandinavian journal of immunology* vol. 90,1 (2019): e12771. doi:10.1111/sji.12771

Naqvi, Ahmad Abu Turab et al. “Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach.” *Biochimica et biophysica acta. Molecular basis of disease* vol. 1866,10 (2020): 165878. doi:10.1016/j.bbadis.2020.165878

O’Connor BD, Van Auwera G. Genomics in the cloud: using Docker, Gatk, and WDL in Terra. 2020.

Oran, Daniel P, and Eric J Topol. “Prevalence of Asymptomatic SARS-CoV-2 Infection : A Narrative Review.” *Annals of internal medicine* vol. 173,5 (2020): 362-367. doi:10.7326/M20-3012

Pagliari, Maria Teresa et al. “Role of ADAMTS13, VWF and F8 genes in deep vein thrombosis.” *PloS one* vol. 16,10 e0258675. 18 Oct. 2021, doi:10.1371/journal.pone.0258675

Pairo-Castineira, Erola et al. “Genetic mechanisms of critical illness in COVID-19.” *Nature* vol. 591,7848 (2021): 92-98. doi:10.1038/s41586-020-03065-y

Perales-Linares, Renzo, and Sonia Navas-Martin. “Toll-like receptor 3 in viral pathogenesis: friend or foe?.” *Immunology* vol. 140,2 (2013): 153-67. doi:10.1111/imm.12143

Petrovszki, Dániel et al. “Penetration of the SARS-CoV-2 Spike Protein across the Blood-Brain Barrier, as Revealed by a Combination of a Human Cell Culture Model System and Optical Biosensing.” *Biomedicines* vol. 10,1 188. 17 Jan. 2022, doi:10.3390/biomedicines10010188

Pyrc, Krzysztof et al. “The novel human coronaviruses NL63 and HKU1.” *Journal of virology* vol. 81,7 (2007): 3051-7. doi:10.1128/JVI.01466-06

Picchiotti, Nicola et al. Post-Mendelian genetic model in COVID-19. Cardiol Cardiovas Med. .2021. 5 (6): 673-694 https://doi.org/10.26502/fccm.92920232

Pierotti, Silvia et al. “Androgen modulation of pro-inflammatory and anti-inflammatory cytokines during preadipocyte differentiation.” *Hormone molecular biology and clinical investigation* vol. 4,1 (2010): 483-8. doi:10.1515/HMBCI.2010.059

Polgreen, Philip M et al. “*CFTR* Heterozygotes Are at Increased Risk of Respiratory Infections: A Population-Based Study.” *Open forum infectious diseases* vol. 5,11 ofy219. 1 Nov. 2018, doi:10.1093/ofid/ofy219

Powazniak, Yanina et al. “VWF and ADAMTS13 behavior in estradiol-treated HUVEC.” *European journal of haematology* vol. 86,2 (2011): 140-7. doi:10.1111/j.1600-0609.2010.01545.x

Pozzilli, Paolo, and Andrea Lenzi. “Commentary: Testosterone, a key hormone in the context of COVID-19 pandemic.” *Metabolism: clinical and experimental* vol. 108 (2020): 154252. doi:10.1016/j.metabol.2020.154252

Purcell, Shaun et al. “PLINK: a tool set for whole-genome association and population-based linkage analyses.” *American journal of human genetics* vol. 81,3 (2007): 559-75. doi:10.1086/519795

Ranjith-Kumar, C T et al. “Effects of single nucleotide polymorphisms on Toll-like receptor 3 activity and expression in cultured cells.” *The Journal of biological chemistry* vol. 282,24 (2007): 17696-705. doi:10.1074/jbc.M700209200

Rastrelli, Giulia et al. “Low testosterone levels predict clinical adverse outcomes in SARS-CoV-2 pneumonia patients.” *Andrology* vol. 9,1 (2021): 88-98. doi:10.1111/andr.12821

Roesch, Erica A et al. “Inflammation in cystic fibrosis: An update.” *Pediatric pulmonology* vol. 53,S3 (2018): S30-S50. doi:10.1002/ppul.24129

Sarantis, Panagiotis et al. “Are cystic fibrosis mutation carriers a potentially highly vulnerable group to COVID-19?.” *Journal of cellular and molecular medicine* vol. 24,22 (2020): 13542-13545. doi:10.1111/jcmm.15941

Sawicki, Stanley G et al. “A contemporary view of coronavirus transcription.” *Journal of virology* vol. 81,1 (2007): 20-9. doi:10.1128/JVI.01358-06

Severe Covid-19 GWAS Group et al. “Genomewide Association Study of Severe Covid-19 with Respiratory Failure.” *The New England journal of medicine* vol. 383,16 (2020): 1522-1534. doi:10.1056/NEJMoa2020283

Shah, Viral S et al. “Relationships among CFTR expression, HCO3- secretion, and host defense may inform gene- and cell-based cystic fibrosis therapies.” *Proceedings of the National Academy of Sciences of the United States of America* vol. 113,19 (2016): 5382-7. doi:10.1073/pnas.1604905113

Sharma, Atul et al. “Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2): a global pandemic and treatment strategies.” *International journal of antimicrobial agents* vol. 56,2 (2020): 106054. doi:10.1016/j.ijantimicag.2020.106054

Sharma, Anshika et al. “COVID-19: A Review on the Novel Coronavirus Disease Evolution, Transmission, Detection, Control and Prevention.” *Viruses* vol. 13,2 202. 29 Jan. 2021, doi:10.3390/v13020202

Shelton, Janie F et al. “Trans-ancestry analysis reveals genetic and nongenetic associations with COVID-19 susceptibility and severity.” *Nature genetics* vol. 53,6 (2021): 801-808. doi:10.1038/s41588-021-00854-7

Schulz, Oliver et al. “Toll-like receptor 3 promotes cross-priming to virus-infected cells.” *Nature* vol. 433,7028 (2005): 887-92. doi:10.1038/nature03326

Sidey-Gibbons, Jenni A M, and Chris J Sidey-Gibbons. “Machine learning in medicine: a practical introduction.” *BMC medical research methodology* vol. 19,1 64. 19 Mar. 2019, doi:10.1186/s12874-019-0681-4

Smith, Everett Clinton, and Mark R Denison. “Coronaviruses as DNA wannabes: a new model for the regulation of RNA virus replication fidelity.” *PLoS pathogens* vol. 9,12 (2013): e1003760. doi:10.1371/journal.ppat.1003760

Song, Zhiqi et al. “From SARS to MERS, Thrusting Coronaviruses into the Spotlight.” *Viruses* vol. 11,1 59. 14 Jan. 2019, doi:10.3390/v11010059

Stanelle-Bertram, Stephanie et al. “CYP19A1 mediates severe SARS-CoV-2 disease outcome in males.” *Cell reports. Medicine* vol. 4,9 (2023): 101152. doi:10.1016/j.xcrm.2023.101152

Su, Shuo et al. “MERS in South Korea and China: a potential outbreak threat?.” *Lancet (London, England)* vol. 385,9985 (2015): 2349-50. doi:10.1016/S0140-6736(15)60859-5

Su, Shuo et al. “Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses.” *Trends in microbiology* vol. 24,6 (2016): 490-502. doi:10.1016/j.tim.2016.03.003

Suresh, Madathilparambil V et al. “TLR3 absence confers increased survival with improved macrophage activity against pneumonia.” *JCI insight* vol. 4,23 e131195. 5 Dec. 2019, doi:10.1172/jci.insight.131195

Suwinski, Pawel et al. “Advancing Personalized Medicine Through the Application of Whole Exome Sequencing and Big Data Analytics.” *Frontiers in genetics* vol. 10 49. 12 Feb. 2019, doi:10.3389/fgene.2019.00049

Tay, Matthew Zirui et al. “The trinity of COVID-19: immunity, inflammation and intervention.” *Nature reviews. Immunology* vol. 20,6 (2020): 363-374. doi:10.1038/s41577-020-0311-8

Tang, Ning et al. “Abnormal coagulation parameters are associated with poor prognosis in patients with novel coronavirus pneumonia.” *Journal of thrombosis and haemostasis : JTH* vol. 18,4 (2020): 844-847. doi:10.1111/jth.14768

Teimouri H, Maali A. Single-nucleotide polymorphisms in host pattern-recognition receptors show association with antiviral responses against SARS-CoV-2, in-silico trial. JoMMID. 2020;8(2):65–70.

The New York Times. Coronavirus Updates: The Illness Now Has a Name, COVID-19. Online: https://www.nytimes.com/2020/02/11/world/asia/coronavirus-china.html (accessed on 17 May 2020)

Totura, Allison L et al. “Toll-Like Receptor 3 Signaling via TRIF Contributes to a Protective Innate Immune Response to Severe Acute Respiratory Syndrome Coronavirus Infection.” *mBio* vol. 6,3 e00638-15. 26 May. 2015, doi:10.1128/mBio.00638-15

Tregouet, David-Alexandre et al. “Specific haplotypes of the P-selectin gene are associated with myocardial infarction.” *Human molecular genetics* vol. 11,17 (2002): 2015-23. doi:10.1093/hmg/11.17.2015

van der Made, Caspar I et al. “Presence of Genetic Variants Among Young Men With Severe COVID-19.” *JAMA* vol. 324,7 (2020): 663-673. doi:10.1001/jama.2020.13719

van der Made, Caspar I et al. “Clinical implications of host genetic variation and susceptibility to severe or critical COVID-19.” *Genome medicine* vol. 14,1 96. 19 Aug. 2022, doi:10.1186/s13073-022-01100-3

Van Vliet, Monique et al. “Hypogonadism, quadriceps weakness, and exercise intolerance in chronic obstructive pulmonary disease.” *American journal of respiratory and critical care medicine* vol. 172,9 (2005): 1105-11. doi:10.1164/rccm.200501-114OC

Wambier, Carlos Gustavo, and Andy Goren. “Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection is likely to be androgen mediated.” *Journal of the American Academy of Dermatology* vol. 83,1 (2020): 308-309. doi:10.1016/j.jaad.2020.04.032

Wang, Chen et al. “A novel coronavirus outbreak of global health concern.” *Lancet (London, England)* vol. 395,10223 (2020): 470-473. doi:10.1016/S0140-6736(20)30185-9

Wang, Yijin et al. “SARS-CoV-2 infection of the liver directly contributes to hepatic impairment in patients with COVID-19.” *Journal of hepatology* vol. 73,4 (2020): 807-816. doi:10.1016/j.jhep.2020.05.002

Wang, Kai et al. “ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.” *Nucleic acids research* vol. 38,16 (2010): e164. doi:10.1093/nar/gkq603

Wang, Leyi et al. “Detection and genetic characterization of deltacoronavirus in pigs, Ohio, USA, 2014.” *Emerging infectious diseases* vol. 20,7 (2014): 1227-30. doi:10.3201/eid2007.140296

Weiss, Susan R, and Sonia Navas-Martin. “Coronavirus pathogenesis and the emerging pathogen severe acute respiratory syndrome coronavirus.” *Microbiology and molecular biology reviews : MMBR* vol. 69,4 (2005): 635-64. doi:10.1128/MMBR.69.4.635-664.2005

WHO R&D Blueprint Novel Coronavirus COVID-19 Therapeutic Trial Synopsis, 2020

Wiersinga, W Joost et al. “Pathophysiology, Transmission, Diagnosis, and Treatment of Coronavirus Disease 2019 (COVID-19): A Review.” *JAMA* vol. 324,8 (2020): 782-793. doi:10.1001/jama.2020.12839

Williams, Frances M K et al. “Self-Reported Symptoms of COVID-19, Including Symptoms Most Predictive of SARS-CoV-2 Infection, Are Heritable.” *Twin research and human genetics : the official journal of the International Society for Twin Studies* vol. 23,6 (2020): 316-321. doi:10.1017/thg.2020.85

Wu, Chaomin et al. “Risk Factors Associated With Acute Respiratory Distress Syndrome and Death in Patients With Coronavirus Disease 2019 Pneumonia in Wuhan, China.” *JAMA internal medicine* vol. 180,7 (2020): 934-943. doi:10.1001/jamainternmed.2020.0994

Xiao, Kangpeng et al. “Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins.” *Nature* vol. 583,7815 (2020): 286-289. doi:10.1038/s41586-020-2313-x

Xu, Hao et al. “High expression of ACE2 receptor of 2019-nCoV on the epithelial cells of oral mucosa.” *International journal of oral science* vol. 12,1 8. 24 Feb. 2020, doi:10.1038/s41368-020-0074-x

Zhang, Sai et al. “Common and rare variant analyses combined with single-cell multiomics reveal cell-type-specific molecular mechanisms of COVID-19 severity.” *medRxiv : the preprint server for health sciences* 2021.06.15.21258703. 21 Jun. 2021, doi:10.1101/2021.06.15.21258703. Preprint.

Zhao, Jincun et al. “Rapid generation of a mouse model for Middle East respiratory syndrome.” *Proceedings of the National Academy of Sciences of the United States of America* vol. 111,13 (2014): 4970-5. doi:10.1073/pnas.1323279111

Zguro, Kristina et al. “Host genetic basis of COVID-19: from methodologies to genes.” *European journal of human genetics : EJHG* vol. 30,8 (2022): 899-907. doi:10.1038/s41431-022-01121-x

Zguro, Kristina et al. “Carriers of *ADAMTS13* Rare Variants Are at High Risk of Life-Threatening COVID-19.” *Viruses* vol. 14,6 1185. 29 May. 2022, doi:10.3390/v14061185

Zhang T, et al. Pangolin Homology Associated with 2019-nCoV. bioRxiv 2020.

Zheng, Ying-Ying et al. “COVID-19 and the cardiovascular system.” *Nature reviews. Cardiology* vol. 17,5 (2020): 259-260. doi:10.1038/s41569-020-0360-5

Zhou, Fei et al. “Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study.” *Lancet (London, England)* vol. 395,10229 (2020): 1054-1062. doi:10.1016/S0140-6736(20)30566-3

Zhu, Na et al. “A Novel Coronavirus from Patients with Pneumonia in China, 2019.” *The New England journal of medicine* vol. 382,8 (2020): 727-733. doi:10.1056/NEJMoa2001017

Zitzmann, Michael, and Eberhard Nieschlag. “The CAG repeat polymorphism within the androgen receptor gene and maleness.” *International journal of andrology* vol. 26,2 (2003): 76-83. doi:10.1046/j.1365-2605.2003.00393.x

Zumla, Alimuddin et al. “Middle East respiratory syndrome.” *Lancet (London, England)* vol. 386,9997 (2015): 995-1007. doi:10.1016/S0140-6736(15)60454-8