University of Siena – Department of Medical Biotechnologies

Doctorate in Genetics, Oncology and Clinical Medicine

(GenOMeC)

XXXIV Cycle (2018-2021)

Coordinator: Prof. Francesca Ariani

Identifying host genetics risk factors for COVID-19 from Exome Sequencing

Scientific disciplinary sector: ING-INF/06 – Electronic and informatics bioengineering

Tutor:

Prof. Simone Furini

PhD Candidate:

Elisa Benetti

ELISA BENETTI

Firmato il 18/10/2021

Academic Year: 2020/2021

cnag
centre nacional d'anàlisi genòmica
centro nacional de análisis genómico

CRG
Centre for Genomic Regulation

Barcelona, October 14th 2021

Review of Elisa Benetti's thesis dissertation "Identifying host genetics risk factors for COVID-19 from Exome Sequencing"

Dear Elisa, Simone and Alessandra,

I have read and reviewed Elisa's thesis with great interest. It is very well structured and written. It clearly shows that Elisa has done a lot of work. Besides learning a lot in the process, she has already published relevant scientific results. It is also remarkable that Elisa, with your supervision, has shown to be able to work under pressure, delivering results in short time when half of the world was on hold.

Although the text is already excellent, I would like to provide some feedback for your consideration. You will see all my comments in the attached Word document (and also copied as a list under this letter). My main suggestions for improvement are the following:

- Expand the introduction on certain topics like COVID-19, ACE2, so that non-experts and future readers can understand
- Explain better WES vs WGS, and how both (or at least WES) is done. A Figure would help.
- Consider adding additional figures (e.g. sequencing WGS/WES, mechanisms giving rise to CNVs, MLPA vs aCGH vs WES?, CNV overlap approach). It might be too challenging to create them from scratch given the short time, but could adapt existing ones.
- Would be good to clearly state the objective/s for each of the results chapters before the short technical and results summary.

Hoping that these suggestions will be useful, I look forward to knowing more about your work on the 21st of October during the online evaluation committee.

Sincerely,

BELTRAN AGULLO SERGI - 36531863C

Firmado digitalmente por BELTRAN AGULLO SERGI - 36531863C
Nombre de reconocimiento (DN): c=ES, serialNumber=IDCES-36531863C, givenName=SERGI, sn=BELTRAN AGULLO, cn=BELTRAN AGULLO SERGI - 36531863C
Fecha: 2021.10.14 17:30:38 +02'00'

Sergi Beltran, PhD
Bioinformatics Unit Head
Centro Nacional de Análisis Genómico (CNAG-CRG)
Barcelona, Spain

baldiri reixac, 4
pcb - tower i, 2nd floor
08028 barcelona

t +34 93 4020542
f +34 93 4037279
www.cnag.crg.eu

ISTITUTO DI RICERCHE
FARMACOLOGICHE
MARIO NEGRI · IRCCS

From: Maddalena Fratelli
Head, Pharmacogenomics Unit
Istituto di Ricerche Farmacologiche Mario Negri IRCCS
Milano

To whom it may concern

I read the thesis entitled: "**Identifying host genetics risk factors for COVID-19 from Exome Sequencing**" written by Elisa Benetti.

The thesis describes an impressive piece of work, carried out in a relatively short time by a collaborative interdisciplinary team, which evidently integrated many different skills to address a complex problem such as the genetic component of the variability of the host response to SARS-CoV-2 virus infection. The results of such an effort are partly described in five papers that have already been published in five reputable journals.

The text is well written. The methods, results and conclusions are clearly and comprehensively described in an easily readable and concise manner. Where necessary, the limitations of the approaches and of the results were adequately highlighted and discussed. As previously remarked, the majority of the results have been published and therefore have been subjected to peer review,

The contribution of the candidate to the whole work should probably be described in more detail, to highlight her competences, that seem to span from the biomedical field to the bioinformatics technicalities.

Yours sincerely

Milano, 14 October 2021

Summary

The quote *"Not everything that can be counted counts and not everything that counts can be counted"*, often attributed to Albert Einstein, expresses in some extent the challenges we are facing when dealing with the human genome. The unprecedent amount of data derived from sequencing experiments forced us to find something that *counts* within an overwhelming number of genetic variants. In the present thesis, we try to assess this issue in the context of Coronavirus disease 2019 (COVID-19), an infectious disease caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). While most infected individuals experience only mild or no symptoms, severe cases can rapidly evolve toward a critical respiratory distress syndrome and multiple organ failure [1]. COVID-19 has demonstrated itself to be a heterogeneous and multifactorial infection having a broad spectrum of clinical presentations influenced by age, gender, comorbidities, ethnic groups, and host genetics, including human leukocyte antigen (*HLA*) genotypes [2]. In this challenging context, our aim was to study host genetic factors associated with COVID-19 severity. A better understanding of the interplay between host genetics and SARS-CoV-2 is, in fact, essential for disease prediction and to support the development of targeted therapies. Several efforts have been done worldwide to discover the genetic determinants of COVID-19 susceptibility, severity, and outcomes. As a matter of fact, COVID-19 represents one of the hot research topic areas for its relevance among the whole community (The COVID-19 Host Genetics Initiative, HGI, and the COVID Human Genetic Effort, HGE, Consortia).

This dissertation presents a novel approach to identify host risk factors predisposing to the disease. The innovation consists in taking into account different aspects of genome variability, from Single Nucleotide Variants (SNVs) to Copy Number Variations (CNVs) through a gene-based approach to represent genetic data. The gene-based Boolean representations were the input features of machine learning models and were tested separately and ultimately all together to improve our ability

to predict COVID-19 outcomes and to identify genes and variants predisposing to severe outcomes. Overall, this method led us to identify some important genetic determinants involved in COVID-19 severity that will be discussed in the final chapters of the thesis.

The first Chapter of this thesis will provide an overview of the background and state of the art technologies to guide the reader in the comprehension of the work. Chapter 2 will provide an exhaustive description of the bioinformatic pipelines, optimization procedures and methods adopted in our work. Chapters 3 and 4 will show our first findings and introduce the reader to the complexity of the study. The effective applications of our novel approach, i.e., the Boolean features and machine learning model, are reported in Chapter 5, 6 and 7. The last chapter of the results, Chapter 8, will discuss the challenges and results of the application of machine learning methods on Boolean features representing copy number variants. The main stages and discoveries of our research will be reported and commented in the Concluding remarks, that end the dissertation on Chapter 9.

1. Introduction

In this chapter, we outline the characteristics of COVID-19, focusing our attention on the role played by host genetics in predisposing to COVID-19 severity. As an adequate method able to represent and explain the complexity of COVID-19 disease is required, the key components of our novel approach, e.g., synthetic representation of genetic data and machine learning models, are described in this chapter. In particular, the choice of Whole Exome Sequencing is contextualized in section 1.2 followed by the description of the variants included in our analyses. A brief overview of the machine learning techniques is illustrated in section 1.6.

1.1 Disentangling complex diseases: the COVID-19 pandemic

The coronavirus disease 2019 (COVID-19, 'CO' stands for corona, 'VI' for virus, and 'D' for disease) pandemic, caused by infections with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), that firstly appeared in December 2019 in Wuhan (China), has resulted in an enormous challenge to the world's healthcare systems. Globally, as of 14th October 2021, there have been 239.007.759 confirmed cases of COVID-19 and 4.871.841 deaths, while in Italy the confirmed cases are attested around 4.707.087 and 131.421 deaths, reported to the World Health Organization (WHO) [3]. As the virus continues to circulate worldwide, the risk of occurrence of new variants, that might have higher infectivity, transmissibility, and virulence, is elevated. Up to now, four SARS-CoV-2 variants of concern (VOCs) have been defined: Alpha (B.1.1.7, first documented in the UK), Beta (B.1.351, first documented in South Africa), Gamma (P.1, first documented in Brazil), and Delta (B.1.617.2, first documented in India) [4].

It is well known now that COVID-19 is characterized by a highly heterogeneous phenotypic presentation. A wide range of symptoms have been reported including fever, cough, difficulty breathing or shortness of breath, fatigue,

loss of taste or smell, sore throat, headache, diarrhoea, muscle or body aches and rush on skin (https://www.who.int/health-topics/coronavirus). Mild cases are defined as patients having, just to mention a few, fever, cough, chest pain, nausea, and body pain while severe and critical cases are those experiencing dyspnoea, respiratory failure and/or septic shock (www.cdc.gov/coronavirus/). While most infected individuals experience only mild or no symptoms, severe cases can rapidly evolve toward a critical respiratory distress syndrome and multiple organ failure, or to persistent disease (long COVID), or death [1], [5]. The risk of severe COVID-19 outcomes is strongly correlated with age, with a doubling in risk every 5 years from the age of 5 years ahead and a strong rise after the age of 65 years [6]. Additionally, other known risk factors are sex, as a male bias in mortality has emerged during the pandemic [7], and the presence of underlying medical conditions like cardiovascular disease, asthma, diabetes, chronic respiratory disease, chronic kidney disease, or cancer [8].

Early in 2020, Italy was the first European country to experience the COVID-19 outbreak with an overall case fatality rate of 7.2%, substantially higher than in China (2.3%) [9]. For this reason, we started investigating the population-specific variation of the coding variants of Angiotensin-converting enzyme 2 (*ACE2*), the SARS-CoV-2 receptor for host cell entry [10]. During the infection, SARS-CoV-2 binds to ACE2 receptor through the Spike glycoprotein (S) and the invasion process is then triggered by host cell proteases (furin, trypsin, TMPRSS2 and cathepsin). When viral RNA enters the host cell, translation of the polyproteins begins [11]. For its role in the virus entry into the host cell, *ACE2* gene was the first target of our study.

Later, we established a consortium, the GEN-COVID Multicenter Study, to study the COVID-19 Host Genetics factors (https://sites.google.com/dbm.unisi.it/gen-covid). In fact, while being a male, increasing age and higher mass index are recognised to be risk factors correlating with disease severity, they do not explain alone all the observed variability among individuals [12]. The interindividual variability in COVID-19 susceptibility and disease severity suggests that a predisposing host genetic background can play a role in the pathogenesis of the disease.

Existing studies suggest that variability in the host genetic constitution, along with immunological features, may modulate the inter-individual and population-scale differences in COVID-19 severity and clinical outcomes [13], [14]. Identifying host-specific genetic factors may provide insight about biological mechanisms leading to disease and consequentially help to support the development of novel treatments. As new virus variants arise, the search for therapies is, in fact, still relevant despite the recent development of vaccines.

Classical studies, such as Genome-Wide Association Studies (GWASs) have been extensively employed to identify some loci associated with COVID susceptibility/protection. As a result, some common polymorphisms in relevant genes have been found in the course of the last two years [15], [16]. However, COVID-19 has demonstrated to be a complex disorder where both common and rare variants contribute to the likelihood of developing a severe form of the disease. Since GWAS studies focus primarily on common variants (MAF>5%), rare variants constitute a missed heritability for this method. Moreover, the variants identified through GWAS explain only a small fraction of trait variability and being mostly non-coding, they make it difficult to interpret the results: follow-up analyses are therefore necessary to identify the relevant genes.

In our effort to untangle COVID-19 complexity, we employed Whole Exome Sequencing (WES) to characterize both common and rare variants as potential contributors to the severe phenotypes. An overview of the strengths and characteristics of this technology is provided in the next section.

1.2 Whole Exome Sequencing (WES)

Advances in Next Generation Sequencing, technologies for massive-parallel DNA sequencing, have resulted in an extraordinary amount of genomic sequence data allowing for a more comprehensive understanding of human genetics. Despite Whole Genome Sequencing (WGS) provides the most extensive analyses of the entire human genome (3 billion base pairs), this approach is not yet considered to offer

sufficiently improved clinical utility with its markedly higher costs compared to exome sequencing and gene panels. A common alternative to WGS is Whole Exome Sequencing (WES) [17], a more cost-effective method that delivers a higher coverage, allowing for detection of variants at lower percentage within the sample (e.g., somatic mutations, mosaics, heteroplasmy). The WES analysis workflow is reported in Figure 1. WES focuses mainly on the protein coding regions of the genome (exons), which encompasses only 3.09% (30Mb) of the latest release of the human reference genome, GRCh38 [18]. Different exome capture kits and providers are available, which primarily differ in their specific genomic target regions covered, size and number of probes [19]. In most cases, the target enrichment strategy includes ~22,000 genes and harbours more than 85% of the variants causing single-gene disorders [19]. For these reasons, WES has been widely and successfully used for the identification of the genetic basis of both Mendelian diseases as well as complex traits.

On average, the WES of a patient generates more than 20,000 variants. The challenge is to determine which of these variants underlie or are responsible for the inherited components of phenotypes by filtering out common variants and prioritizing candidate variants [20]. There are different classes of genetic variations such as Single Nucleotide Variants (SNVs), small insertion and deletion variants (INDELs), Copy Number Variants (CNVs), and large Structural Variants (SVs). While SNVs and INDELs are routinely detectable by WES variant calling, the ability to detect CNVs and SVs has only recently emerged and presents considerable challenges.

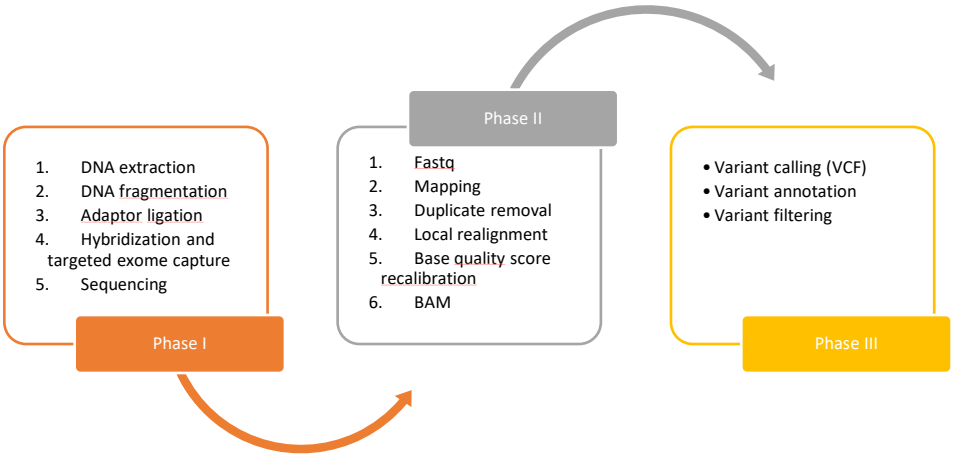


Figure 1. Whole Exome Sequencing workflow

1.3 Single nucleotide variants (SNVs) and small insertion and deletion variants (INDELs)

Single nucleotide variants (SNVs) are among the most frequent and widespread alterations in the genome [21]. The vast majority of these changes are functionally neutral; however, some variants produce dramatic phenotype and may cause diseases as a consequence. While not as common as SNVs, INDELs are widely spread in the genome. They are a type of genetic variation in which a specific nucleotide sequence is inserted or deleted. A comprehensive summary of the types of sequence variation is reported in Figure 2, taken from (https://m.ensembl.org/).

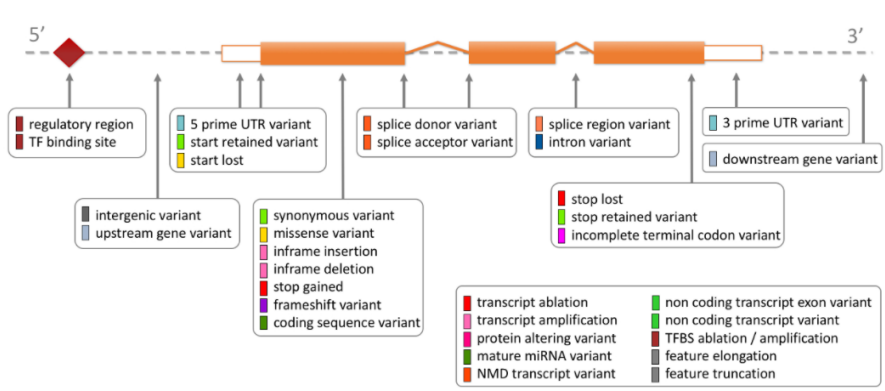


Figure 2. Diagram showing the location of sequence variation. (Picture taken from https://m.ensembl.org/)

A variant may fall within the coding region of genes (synonymous variants, missense variants, frameshift caused by INDELs, in frame variants, stop gained), non-coding regions of genes (e.g., 5'UTR variants, 3'UTR variants), in the boundaries between exons and introns (splice variants), or in the intergenic regions between genes (intergenic variant, upstream and downstream gene variants). SNVs within a coding sequence do not necessarily change the amino acid sequence of the protein, due to degeneracy of the genetic code (synonymous variants).

Nearly half of the known inherited disease mutations are non-synonymous SNVs (nsSNVs) [22], which by causing an amino acid change can destroy the function of the encoded proteins. The high number of detected variants make impossible to investigate the functional effect of every nsSNVs experimentally. Thus, the interpretation of genetic variants remains an enormous challenge and further development of methods to prioritize variants that are clinically relevant is essential to maximize the utility of sequencing data. As a consequence, variants' annotation – which assigns functional information to DNA variants – is a key step in any bioinformatic pipeline for the analysis of WES data. Multiple tools have been developed for predicting deleteriousness of genetic variants such as SIFT, MutationTaster, PhyloP, FATHMM, MutationAssessor, POLYPHEN2, CADD,

GERP++. These programs rely on different methods and provide a score that measures how likely an nsSNVs is to be deleterious, along with its binary prediction. Some tools measure evolutionary sequence conservation (e.g. SIFT, PhyloP and MutationAssessor) using mathematical operations. Other tools evaluate the impact of variants on protein structure and function using physical and comparative factors (e.g., PolyPhen-2), classifying variants according to Bayesian methods. Another class of tools is represented by the ones that predict the overall pathogenic potential of a variant integrating a number of genomic information, such as sequence context, epigenetic measurements, gene model annotation, and using a machine learning approach to categorize variants as benign or deleterious (e.g., CADD). Despite the important guidance on variant interpretation provided by these tools, the predictions can vary greatly when applied to the same variants [23], suggesting that further improvements are still needed. In particular, the low specificity of the current tools entails a high rate of false positive predictions, which complicate the identification of causative variants.

1.4 Copy number variations (CNVs)

The routine use of WES is generating a great amount of inconclusive data. As a matter of fact, most patients with a suspected genetic condition are left undiagnosed even after a thorough analysis of rare coding SNVs and INDELs [24]. This can occur for various reasons, including the lack of knowledge of genes leading to a focused analysis of only known disease genes, but it can also be due to different type of variation not routinely detectable by WES analysis pipelines, such as structural variations.

The term “structural variations” comprises microscopic and sub-microscopic variants which include duplications and deletions, collectively called copy-number variants or copy-number polymorphisms, as well as insertions, inversions and translocations [18]. These variations may impact the dosage or the regulation of one or more genes or generate somatic genome instability and age-dependent diseases.

Deletions and duplications are a type of structural variation referred to as copy number variations (CNVs) involving copy number changes of DNA fragments typically longer than 1 Kb [25]. The 1000 bp threshold derives from earlier studies based on microarray methods but currently the size of CNVs is defined from 50bp to several Mbs after the application of sequencing technologies [26].

CNVs are common features of the human genome and account for more inter-individual variation than do single-nucleotide variants. Their impact ranges from no obvious effect on common variability of physiological traits, to substantial contribution to common and rare diseases susceptibility [27]. Pathogenic CNVs have been found to cause Mendelian disorders [28] or to be associated with complex multifactorial diseases, including cancer [29], cardiovascular [30] and neurodevelopmental disorders [31], and to contribute to susceptibility to infectious diseases [32].

1.4.1 Mechanisms of CNVs formation

There are four major mechanisms giving rise to CNVs, reported in Figure 3. Two recombination-based mechanisms such as NAHR (Non-Allelic Homologous Recombination) between repeat sequence and NHEJ (Non-Homologous End-Joining) have been linked to genomic rearrangements and the formation of CNVs together with retrotransposition and a replication-based mechanism termed FoSTeS (fork stalling and template switching) [33]. NAHR appears to be the predominant pathway underlying recurrent rearrangements of the genome. It is caused by the alignment and the following crossover between two nonallelic (i.e., paralogous) DNA sequence repeats sharing high similarity to each other. NAHR can take place in meiosis where it results in unequal crossing over leading to constitutional genomic rearrangements, but it can also occur in mitosis resulting in mosaic populations of somatic cells carrying copy number variations. NHEJ is responsible for many of the nonrecurrent rearrangements [28]. This mechanism is used by human cells to repair double strand breaks and can result in several nucleotides loss or addition at the join

point [33]. The FoSTeS model can also account for Complex Genomic Rearrangements and CNVs. According to this mechanism, the DNA replication fork can stall, the lagging strand separates from the original template and switches to another replication fork and restarts DNA synthesis on the new fork by priming it via the microhomology between the switched template site and the original fork [33]. Depending on whether the new fork is located downstream or upstream of the original fork, the template switching results in either a deletion or a duplication. Moreover, depending on the orientation of the replication fork, the erroneously inserted fragment could be in direct or inverted orientation compared to its original position. This whole procedure can take place multiple times in series resulting in complex rearrangements. Even if the vast majority of gene duplications results in a new copy located adjacent to the original gene, a substantial number of new duplicates are inserted far from the original locus in humans [34]. In this case, the underlying mechanism is the retrotransposition in which a mRNA transcript is reverse-transcribed and reinserted into a random location in the genome, yielding a new intron-less gene copy.

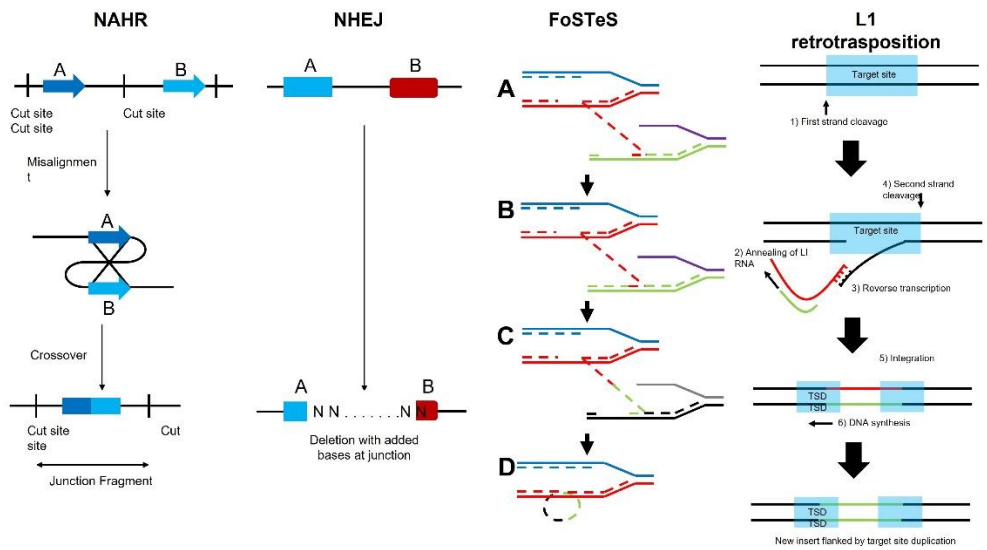


Figure 3. Mechanisms of CNVs formation.

1.4.2 CNV detection methods

CNV detection methods can be divided into two major categories: locus-specific detection, which requires prior knowledge of the region of interest; and genome-wide detection, which allows CNVs detection across the whole genome or a considerable part of it [35]. Among locus-specific methods are multiplex ligation-dependent probe amplification (MLPA), quantitative polymerase chain reaction (qPCR) and Fluorescence in situ hybridization (FISH). Even if the locus-specific techniques are considered to be the most reliable, they present some drawbacks, above all the fact that in most cases which region need to be tested for CNVs is not known a priori. Therefore, locus specific methods are often employed to validate selected findings of genome-wide methods. Among genome-wide techniques are array comparative genomic hybridization (aCGH), SNP-arrays and WES or WGS. With these techniques, it is not possible to achieve the same reliability of locus-specific methods, but they often provide an overview of many potential events.

The gold standard for CNV detection in clinics are MLPA and aCGH. MLPA is a targeted PCR-based method that simultaneously analyses multiple genomic regions of interest to detect abnormal copy numbers at an exon-level resolution. It works by quantifying probes that hybridize to genomic DNA and are amplified by PCR. The products are then separated by capillary electrophoresis. Relative amounts of probe amplification products reflect the relative copy number of target sequence [36]. aCGH is based on the principle of comparative hybridization of two labelled samples (test and reference) to a set of hybridization targets. The resulting fluorescent ratio is then measured, converted to a log2 ratio, and used as a proxy for copy number. An increased log2 ratio corresponds to a gain in copy number in the test compared with the reference; conversely, a decrease indicates a loss in copy number [26]. Detection of a CNVs typically requires a signal from at least 3 to 10 consecutives probes. aCGH can reliably call large CNVs (in the order of megabases) but shows poor performances when dealing with small CNVs affecting only one or a few small exons, due to its low resolution (approximately 10~25 kbp) [37]. Employing WES to predict CNVs could extend the diagnostic yield and increase the utility of these

previously unused data, saving time and reducing costs of laboratories, while creating a more comprehensive snapshot of genomic variation with a single assay [24].

1.4.3 CNV prediction from Sequencing data

Reliable CNV calls from sequencing data presents considerable challenges and depends on high depth and uniformity of coverage across targets. Additionally, no accepted standard protocols or quality control measures are available so far [38], [39]. Limitations of this approach arise from the differences in probe hybridization and efficiency, which introduce bias and noise affecting the uniformity and consistency of coverage across all target sites. A robust bioinformatics approach is required to deal with the size and complexity of the data. Many tools have been developed to detect gains or losses of genetic information from sequencing data and rely essentially on 4 different strategies: Read Depth (RD), Paired-end mapping (PEM), Split read (SR), and Assembly (AS) (Figure 4).

Read Depth (RD) methods are based on the hypothesis that there is a correlation between depth of coverage of a genomic region and the copy number of the region. These tools compare the number of reads mapping to a chromosome window with its expectation under a statistical model. Deviations from this expectation are indicative of CNV calls. Limitations of this method are the need of high coverage for high resolution, deletions are detected more easily than duplications and repeats and GC content might introduce artefacts.

Orientation and Distance of Paired-end read mapping (PEM) is based on the distances between a pair of paired-end reads through discordantly mapped reads. A discordant mapping is produced if the distance between two ends of a read pair is significantly different from the average insert size. This approach has the potential to find any type of structural variant (SV) and not only deletions and duplications.

Split reads (SR) method uses reads from paired-end sequencing where only one read of the pair has a reliable mapping while the other one fails to map to the

genome either completely or partially. The unmapped reads are a potential source of breakpoints at the single base pair level.

Unlike the RD, PEM and SR approaches that first align sequencing reads to a known reference genome before the detection of CNVs, in the Assembly (AS) approach contigs are reconstructed from short reads by linking overlapping reads. Genomic regions with discordant copy numbers are detected by comparing the assembled contigs to the reference genome [40], [41].

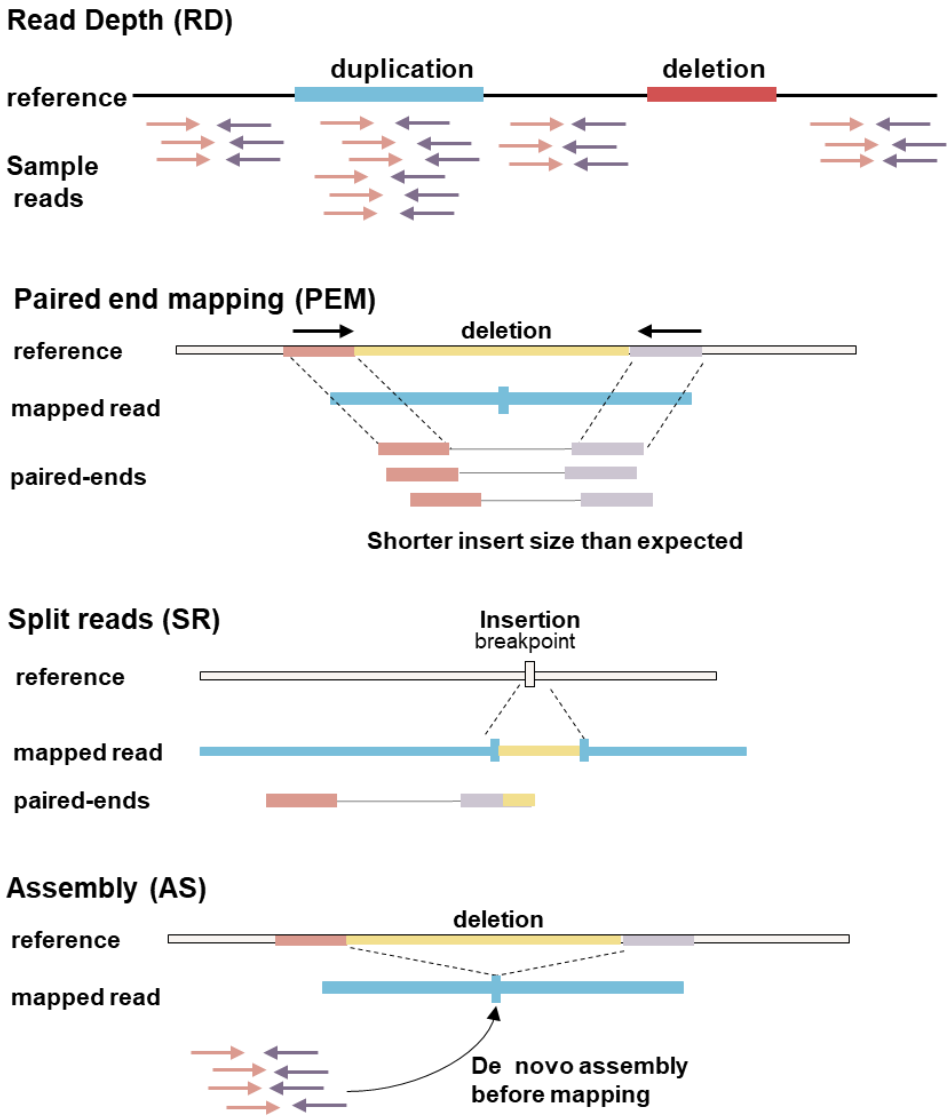


Figure 4. Different strategies of CNV prediction from sequencing data.

When dealing with WES data, the best approach to detect CNVs is through RD based methods, due to the improved sequencing technologies and at the same time the reduced costs which lead to higher coverage data. Like aCGH, the ratio of read counts between a test and a reference sample is preferable than a single-sample analysis in order to correct for the usually broad variability in capture efficiency across exons. PE, SR and AS approaches are instead not suitable for identifying

CNVs from WES data, as exome relies on short and discontinuous exonic regions across the genome [40]. RD-based approaches follow a three-step procedure: mapping, normalization, and estimation of copy numbers. In the mapping stage, short reads are aligned to the reference genome and the read depth is computed according to the number of mapped reads in predefined windows. The second step consists of normalization and correction of potential biases in read depth mainly caused by GC contents and repetitive regions. Lastly, copy number along the chromosomes are estimated to determine deletions or duplications [42]. Currently, a high level of sensitivity can be achieved with these CNV detection tools, but at the cost of low specificity, which increases the workload in interpretation and annotation of CNVs [43].

CNV calls from exomes in this thesis were generated using two RD-based tools which are widely used for this purpose in the field, ExomeDepth [44], and CoNIFER [45]. ExomeDepth and CoNIFER use different statistical models for CNV calling. ExomeDepth is based on Hidden Markov Models and uses a robust beta-binomial model for the modelling. This tool uses a cohort of samples for normalization. An aggregate reference set is created selecting the most suitable control set for each exome by using read count data. This optimized reference set is built in order to maximize the power to detect CNVs [44]. CoNIFER, instead, performs the unsupervised decomposition of the signal using principal component analysis (singular value decomposition). It is based on the assumption that the main source of variability is due to stochastic noise and not to real events. For this reason, the developers of CoNIFER suggest cleaning the cohort of stochastic noise using SVD-based normalization [45].

1.4.4 Challenges associated with detecting CNVs from sequencing data

The advent of high-throughput sequencing technologies is transforming our ability to detecting CNVs. Mainly due to the decreasing cost of sequencing and the increase of high-coverage data, RD-based methods have recently become a major

approach to estimate copy numbers from WES data, where deletions or duplications are identified as decrease or increase of RC across multiple consecutive windows.

RD approach relies on the assumption that the sequencing is uniform, i.e., the coverage follows the Poisson distribution and the number of reads mapped to a region is proportional to the number of copies. A genomic region that has been deleted (duplicated) will have less (more) reads mapping to it than a region not deleted (duplicated). However, the uniformity of coverage across targets might be affected regardless of the copy number of the region, resulting in false positive calls [38]. The main bias against uniform distribution of reads in WES is the capture itself, along with the fact that information is available only on discrete regions. Other biases associated with the sequencing technology exist, including short read lengths, GC-content and mappability.

The percentage of guanine and cytosine in a genomic region varies markedly along the genome and has been found in several studies to influence coverage on many sequencing platforms especially when the GC content is very high or low [46]–[48].

The mappability bias, instead, arises during the alignment step, when a huge number of short reads map to multiple positions in the presence of repetitive regions in the reference genome. (Low mappability regions show large read count overdispersion). As a result, under/over sampled regions caused by biases in sequencing depth other than changes in copy number, affect our ability to detect true deletions/duplications. In order to reduce the effect of these causes of variation and make data comparable within and between samples, Read Counts need to be normalized.

1.5 Mapping genetic variants to gene-based Boolean features

Sequencing-based approaches have been applied in the attempt to identify rare genetic determinants for COVID-19, as they can be associated especially with extreme clinical presentations. With this method, some rare families were identified

with a Mendelian form of inheritance [13], [49]. However, these patients represent only a small fraction of those severely affected by COVID-19. As also common variants might play a role in the contribution to the severe phenotype, we wanted to assess their likely different impact within the same model. In this dissertation, we present a novel approach to consider all genome variability, including variants found at any frequency within a population. Moreover, we were the first, to our knowledge, to evaluate the potential impact of different type of variants, i.e., CNVs, to the overall complex pathogenesis of COVID-19. In 2020, we started to investigate how common variants may combine with rare variants to determine COVID-19 severity in WES data using a first small cohort of hospitalized patients. This pilot analysis revealed that the combination of rare and common variants could potentially impact clinical outcome. In particular, common variants in susceptibility genes may represent the favourable background in which additional host private mutations may determine disease progression [50]. This hypothesis was in line with our previous suggestion that both polymorphic and rare variants in *ACE2* gene, may affect infectivity and partially explain the observed inter-individual clinical variability [10].

Starting from these preliminary results, we aimed to further refine our analysis. In the proposed model, SNVs and CNVs predicted from WES data were converted to gene-based Boolean features, as described in section 2.5. This helped us, on one side, to reduce the dimensionality of the problem (being the number of input features orders of magnitude higher than the number of patients), and, on the other side, to analyse various sources of information within the same model. These Boolean representations were the input features for our analyses to detect the genetics basis of COVID-19 severity.

1.6 Machine Learning

The main idea of Machine Learning (ML) algorithms is to automatically *learn* relevant information from data. Since their performance generally improves as the number of samples increases, the availability of massive genomic datasets has led

to the exponential application of ML techniques in the classification/clustering tasks related to many biological and medical fields.

A classical distinction of ML algorithms is among supervised learning (SL) and unsupervised learning (UL). The difference between these two classes of algorithms is that SL is based on the existence of a dataset, training-set, where the relationship between input features and target variable is known, which instead is not needed in UL. The goal of SL is to take advantage of the training-set for learning a function that best approximates the general relationship between input features and target variable. UL, on the other hand, does not have explicitly labelled outputs, and its goal is to deduce the natural structure presents within a dataset [51].

The effectiveness of ML algorithms depends almost entirely on the particularities of the problem in relation to the available dataset [52]. Usually, SL models are divided into classification or regression problems. In classification problems, labelled data are used to make prediction among a limited set of restricted classes. In these cases, the output variable must be categorical. Instead, in regression models, the target variable is continuous and consequently the goal of the model is to map input features to a continuous output.

Common algorithms in SL include logistic regression (LR), decision tree (DT), support vector machines (SVM), neural networks (NN), and random forests (RF) [52]. K-means clustering and principal component analysis (PCA) are instead some of the most common algorithms for UL.

One of the most critical issue in ML is overfitting. Overfitting, or high variance, refers to learning a function that tries *too hard* to fit the target variable in the training set. As a result, the function fails to generalize to new data points. This can happen when dealing with an overcomplex model (with too many parameters) or when there are too many features compared to the number of observations. Two common strategies to avoid overfitting are to reduce the number of features or to use a regularization technique. Various types of regularization techniques are available,

and they usually operate by adding a penalty term that discourages high values for the model parameters that are not strongly correlated with the output [53].

1.6.1 Application to severity prediction and gene discovery in COVID-19

In our attempt to study host genetic risk factor for COVID-19, we aimed to predict the severity of COVID-19 using information extracted from WES, and at the same time to identify the most relevant genes involved in the classification. For this reason, in this thesis we adopted the Least Absolute Shrinkage and Selection Operator (LASSO) logistic regression that provides a feature selection method within the binary classification tasks (mild *vs* severe) able to enforce both the sparsity and the interpretability of the results [54]. In fact, the coefficients of the LR model are directly linked to the importance of the corresponding features, and LASSO regularization shrinks close to zero the coefficients of features that are not relevant in predicting the response, reducing overfitting and giving direct interpretability of the model predictions in terms of few features importance. In this classification task, the positive weights of the LASSO LR reflect a susceptible behaviour of the features (*i.e.*, genes) to the target COVID-19 severity, whereas the negative weights reflect a protective action of the feature.

As already mentioned in section 1.6, the input features of LASSO logistic regression were the gene-based Boolean representations developed to map the genetic variability into a set of informative features. The decision to move from simple genetic variants to Boolean representations at gene level is due to the necessity of reducing the number of features and at to increase the interpretability of the biological meaning of the extracted features. The target variable, instead, was the COVID-19 severity (severe cases vs mild subjects). The assessment of COVID-19 clinical category is described in the next chapter.

2. Methods

This chapter describes the bioinformatic pipelines used for variants' calling and for the detection of CNVs from WES data. The high dimensionality of the features extracted by these bioinformatic pipelines prevents the application of standard statistical methods for the identification of relevant associations. In order to reduce the dimensionality of the problem and, at the same time, to include prior knowledge into the analysis pipeline, the information extracted from WES were converted into Boolean features. The methods adopted for this conversion are described in section 2.5.1 for SNVs and 2.5.2 for CNVs. The Boolean features defined here will form the bases for applications of Machine Learning models to COVID-19 in the following chapters.

2.1 The GEN-COVID Biobank

The GEN-COVID Multicenter Study involves a network of Italian hospitals and healthcare facilities with the aim to collect and organize biological samples and clinical data along with patient-level phenotypic and genotypic data. To globally share samples and data among COVID-19 researchers, a GEN-COVID Biobank (GCB) and a GEN-COVID Patient Registry (GCPR) were established using already existing biobanking and patient registry infrastructures. For each patient, basic demographic information (sex, age and ethnicity) together with family history, (pre-existing) chronic conditions, and SARS-CoV-2 related symptoms were collected via an extensive clinical questionnaire.

The study protocol also provides access to patients' medical records and continual clinical data updating in order to secure continuity for patient follow-up. The COVID-19 severity was assessed using a slightly modified version of the World Health Organization COVID-19 Outcome Scale [55] as coded into the following seven categories:

-1. resistors to infection (those who despite significant exposure to the virus remain negative)

0. not hospitalized

1. hospitalized, without oxygen support

2. hospitalized, receiving low-flow supplemental oxygen

3. hospitalized, receiving continuous positive airway pressure (CPAP) or bilevel positive airway pressure (BiPAP) ventilation

4. hospitalized receiving invasive mechanical ventilation; and

5. deceased

A total of 2262 samples taken from the GEN-COVID consortium are analysed in this dissertation for the CNVs analysis (Chapter 8). Subsets of this cohort, taken at different time points in the enrolling process, are analysed in the studies presented in Chapter 4,5,6,7. The mean age of the entire cohort is 60.7 years (range 18–99). The cohort is predominantly male (58.6%) with a mean age of 60.9 years (range 18–99); the mean age of the females is 60.4 years (range 18–98) (Table 1). About 31.1% of the cohort has at least one comorbidity. The overall case-fatality rate is 6.8% (155) deaths among 2262 enrolled subjects with a mean age of 77 years (range 39-98). Regarding the ethnicity, the cohort is composed of 2101 White (92.88%,), 52 Hispanic (2.03%), 27 Black (1.19%), and 31 Asian (1.37%) patients (Table 1). Data on ethnicity and clinical category is not available for 15 out of 2262 patients.

Table 1. Cohort characteristics

No. of subjects	**2262**
Mean age (range)	60.7 (18–99)
Gender no. (%)	
Male	1326 (58,6%)
Female	936 (41.4%)
Ethnicity no. (%)	
White	2101 (92.88%)

Hispanic	52 (2.03%)
Black	27 (1.19%)
Asian	31 (1.37%)
Clinical category no. (%)	
Deceased (group 5)	155 (6.8%)
Hospitalized intubated (group 4)	143 (6.3%)
Hospitalized CPAP/BiPAP (group 3)	470 (20.8%)
Hospitalized with oxygen support (group 2)	704 (31.1%)
Hospitalized w/o oxygen support (group 1)	273 (12.1%)
Not hospitalized oligo/asymptomatic (group 0)	470 (20.8%)
Resistors to infection (group -1)	35 (1.5%)

2.2 Sequencing

Whole exome sequencing of 2262 SARS-CoV-2-infected participants from the Italian GEN-COVID cohort was performed using the Illumina NovaSeq6000 System (Illumina, San Diego, CA, USA). Library preparation was performed using the Illumina Exome Panel 45 Mb (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). All samples were aligned to the GRCh38 human genome assembly using BWA mem v0.7.17. SAMtools v1.7 was used to sort and index BAM files. Variant calling was performed according to the GATK best practice guidelines. Annotation was performed with ANNOVAR and VEP.

2.3 Normalization

For CNV prediction, BAM files underwent a series of preparation steps before read depth calculation. These steps included removal of duplicated sequences and

removal of sequences with low mapping quality (MQ). Through Picard's Mark Duplicates tool, technical and optical duplicates were removed. The main reason for removing duplicates is to mitigate the effects of PCR amplification introduced during library construction. Optical duplicates are instead removed because they result from a single amplification cluster and are incorrectly detected as multiple clusters by the optical sensor. The mappability issue was addressed by removing reads with low MQ score (MQ < 20), which usually fall in repetitive regions of the reference genome or have low base quality, which was done using *samtools* [56]. This step was not necessary for ExomeDepth since it is already included in its pipeline by default. To reduce the GC bias, regions with %GC content higher than 80 were removed from the Exome kit by *bedtools* [57].

2.4 CNV detection pipeline

2.4.1 CoNIFER

CoNIFER v.0.2.2 was run with default settings. BAM files were used to calculate RPKM values. RPKM values were then transformed into standardized z-scores (termed ZRPKM values) based on the mean and standard deviation across all analysed WES. CNV detection from WES data was performed separately by sex and by chromosomes (autosomal *vs* sexual). Mitochondrial DNA (mtDNA) and chrY were excluded from the analysis because the number of probes covering these regions were fewer than samples in the analysis. To reduce heterogeneity, the analyses were firstly performed separately by batch of sequencing runs for a total of twelve batches of around 200 samples each.

To assess the quality control of exomes, it is common practice to examine the standard deviation of samples and to remove those samples with extremely high values. As the standard deviation of all SVD-ZRPKM values for each individual was poor for most of samples, the analysis was repeated with only two batches regardless of the sequencing run. The percentage of samples with low standard deviation (< 0.6)

increased from 22% to 46% when analysing two batches *vs* twelve batches and went up to 87.7% when looking for standard deviation below 0.7. For this reason, the following analyses were performed using two batches of 1000 samples. After setting the number of batches, the number of singular values decomposition (svd) to be removed was chosen for each batch according to the inflection point of the generated scree plot, as suggested by Krumm et. *al* (Figure 5). Nine components were removed for both batches. We set the discovery threshold at -1.5 SVD-ZRPKM for deletions and +1.5 for duplications, and required at least three exome probes to exceed the threshold.

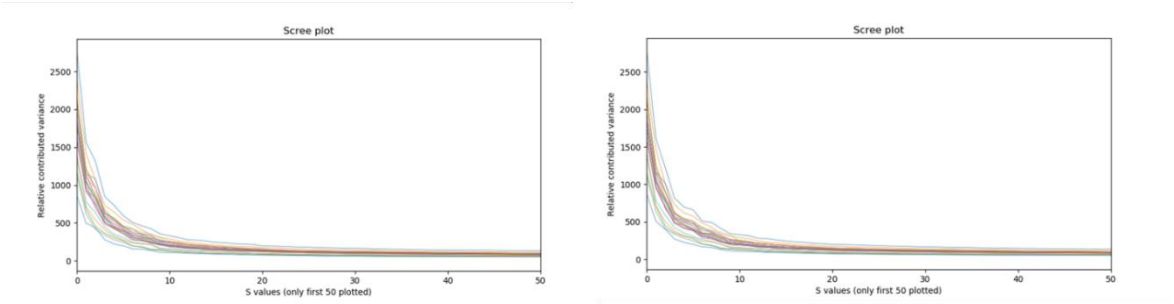


Figure 5. Scree plot of the two batches of samples. 9 svd were removed from both batches.

2.4.2 ExomeDepth

ExomeDepth v1.1.15 was used with default settings. The analysis was performed on autosomal and gonosomal chromosomes separately for males and females, as recommended. Read count data was computed from BAM files and stored into GRanges objects from R. The correlation coefficient between test sample and reference control set was checked to estimate the quality. As it is highly recommended, this value should be above 0.97 or the calling would be less reliable (i.e., most likely a high false positive rate). Most of the samples satisfied this requirement. (Figure 6) ExomeDepth typically selected 10-20 experiments as background.

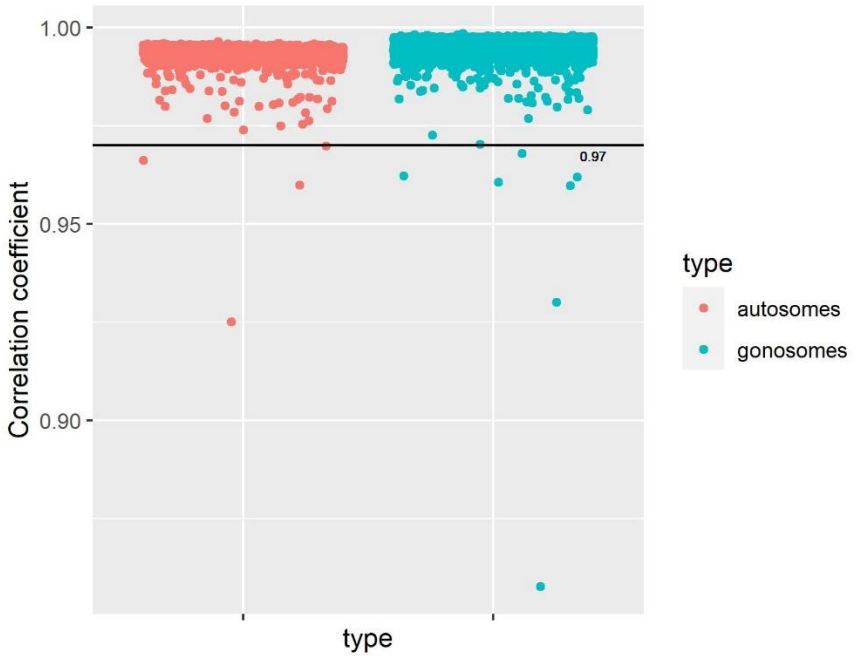


Figure 6. Correlation coefficient between test and reference set separately for autosomes and gonosomes. Most samples have an optimal correlation coefficient above 0.97.

2.4.3 CNV intersection

CNVs predicted with CoNIFER and ExomeDepth were intersected to get a set of more reliable calls. A non-reciprocal overlap was considered, i.e., the two events did not require to have the same minimum overlap. The following commands were used to perform the intersection (Bedtools v2.30).

bedtools intersect -a conifer.calls -b exomedepth.calls -wao -F 0.50

bedtools intersect -a conifer.calls -b exomedepth.calls -wao -f 0.50

Where:

- -*wao* option writes the original A (CoNIFER) and B (ExomeDepth) entries plus the number of base pairs of overlap between the two tools. However, A features w/o overlap are also reported with a NULL B feature and overlap = 0.

- -*F* option is the minimum overlap required as a fraction of B (exomedepth calls). This was chosen because CoNIFER generally detects longer but fewer CNVs than ExomeDepth and we wanted to check if CoNIFER CNVs encompass several ExomeDepth events.

- -*f* option is the minimum overlap required as a fraction of A (conifer calls). We also performed this intersection to take into account also the (few) CNVs which are longer in ExomeDepth than CoNIFER.

Finally, the union of these two sets of overlapped calls was made.

2.5 Definition of the Boolean features

WES data were converted in a binary mode on a gene-by-gene basis. Several types of Boolean representations were implemented to try to cover all the possible genome variation and are described in detail in the next sections. The Boolean representation were then used as input features for LASSO logistic regression models.

2.5.1 Boolean representations of SNVs and INDELs

SNVs and INDELs from WES experiments were collapsed at gene-level and codified into 13 sets of Boolean features. The full list of Boolean representations of SNVs and INDELs is reported in Table 3.

Common variants where the reference is the minor allele were switched. Firstly, any variant not impacting on the protein sequence was discarded. In particular, these were the skipped categories: 'downstream', 'intergenic', 'ncRNA_exonic', 'ncRNA_intronic', 'ncRNA_splicing', 'upstream', 'UTR3', 'UTR5', 'ncRNA_exonic;splicing', 'upstream;downstream', 'upstream;downstream',

'UTR5;UTR3', 'intronic' except for variants classified as pathogenic in Clinvar. 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% ≤ MAF <1%
- low-frequency, 1% ≤ MAF < 5%
- common, MAF ≥ 5%.

Non-Finnish European (NFE) was used as a reference population. SNVs 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.

The rational of the subdivisions based on frequency is to give the appropriate weight to the single variants. A polymorphism is expected to have less impact on the phenotype in respect to a rare variant. More generally, as the frequency decreases, the impact on the protein function is expected to increase. By putting variants of any frequency together in the model, the underlying weight and impact is lost. Ultra-rare and rare variants were divided as the ultra-rare ones, being private, are more likely to have a higher impact for that specific patient,

For the ultra-rare variants, 3 alternative Boolean representations were defined, which were 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. The rational for isolating X chromosomes from autosomal

ones is to preserve the difference between females and males. As males have only one copy of the X chromosome, their AD model for genes on chrX would have a different meaning in respect to genes on autosomes.

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.

Lastly, as common poly-amino acid repeat polymorphisms are usually missed in the classical analysis, such as GWAS analysis (that focus on common bi-allelic polymorphisms), we wanted to test their role in determining COVID-19 clinical severity. Genes with repeated regions were considered in the Boolean of poly-amino acids triplet repeats (C_PR). A total of 40 genes with 43 triplet repeat regions were identified in UniProtKB. For any of these genes two features were defined, Dij and Iij, with Dij equal to 1 if for the i-th patient the j-th gene presented a deletion in the region characterized by repeated triplets, 0 otherwise, and being Iij equal to 1 if for the i-th patient the j-th gene has a repeated region longer than the reference (insertion), 0 otherwise.

Table 2. Boolean representations of SNVs and INDELs

Representations		Boolean categories	
		1	0
UR_AD	Ultra-rare variants (dominant)	At least one variant (MAF < 1/1000)	No this type of variants
UR_AR	Ultra-rare variants (recessive)	At least 2 variants (MAF < 1/1000)	No this type of variants
UR_X	Ultra-rare variants on the X chr genes (X-linked inheritance)	At least one variant (MAF < 1/1000)	No this type of variants
R_AD	Rare variants (dominant)	At least one variant (MAF between 1/100 and 1/1000)	No this type of variants
R_AR	Rare variants (recessive)	At least 2 variants (MAF between 1/100 and 1/1000)	No this type of variants
R_X	Rare variants on the X chr genes (X-linked inheritance)	At least one variant (MAF between 1/100 and 1/1000)	No this type of variants
LF_AD	Low-frequency variants (dominant)	At least one variant (MAF between 5/100 and 1/100) (If more than one coding low-frequency variant impacts in that gene, different combinations - unique-are represented separately)	No this type of variants
LF_AR	Low-frequency variants (recessive)	Variant or variant combination as at LF_AD, in homozygosity (MAF between 5/100 and 1/100)	No this type of variants
LF_X	Low-frequency variants on the X chr genes (X-linked inheritance)	At least one variant (MAF between 5/100 and 1/100) (If more than one coding low-frequency variant impacts in that gene, different combinations -	No this type of variants

		unique-are represented separately)	
C_AD	Common variants (dominant)	At least one variant (MAF > 5/100) (If more than one coding low-frequency variant impacts in that gene, different combinations - unique-are represented separately)	No this type of variants
C_AR	Common variants (recessive)	Variant or variant combination as at C-AD, in homozygosity (MAF > 5/100)	No this type of variants
C_X	Common variants on the X chr genes (X-linked inheritance)	At least one variant (MAF > 5/100) (If more than one coding low-frequency variant impacts in that gene, different combinations - unique-are represented separately)	No this type of variants
C_PR	Common deletion or insertion in genes with repeated regions	Ins = 1 if longer than reference, Del = 1 if shorter than reference	No this type of variants

2.5.2 Boolean representations of copy number variants

The overlapped CNVs between CoNIFER and ExomeDepth were filtered for Bayes Factor (BF), a quality indicator provided by ExomeDepth. BF measures the CNV confidence and depends upon signals arising from a series of contiguous probes. Shorter CNVs detected by fewer probes result with low BF values, while longer CNVs detected by more probes have higher BF values. While it is difficult to assign

an ideal BF threshold and considering that short exons are penalized, we assumed 10 as the best value that minimizes false positive calling rate and maximizes CNV calling number. This threshold was calculated by looking at the lowest BF associated with a confirmed CNV, e.g., with a predicted CNV detected also with another technique. By setting this threshold we preferred to identify a lower number of short CNVs with higher confidence.

To build a gene-based Boolean for CNVs, each event was split by genes spanning through its length, after annotation with AnnotSV [58]. For any of these genes, a Copy Number (CN) value was assigned based on the ratio between observed and expected reads. In particular, a CN-value of 0 indicates homozygous deletion, 1 heterozygous deletion, 2 neutral, 3 heterozygous duplication, 4 homozygous/double duplication, and 5 for duplication values above 4 (Table 2). Three alternative Boolean representations of CNVs were defined. In the representation named CV (Copy Variation), each feature was set equal to 1 if the corresponding gene presented any copy number alteration, and 0 otherwise. The rational of this representation is to capture any possible genes whose copy alterations might have a functional effect on predisposition to COVID-19 severity. Two alternative representations considering only deletion or insertions were defined. In the representation named D (Deletion), each feature was set equal to 1 if the corresponding gene presented either a homozygous or heterozygous deletion (CN-value of 0 and 1) and 0 otherwise. While in the representation named I (Insertion), each feature was set equal to 1 if the corresponding gene presented any duplication (CNV-value of 3,4 and 5) and 0 otherwise. The rational of these representations is to investigate correlations between COVID-19 severity and respectively deletions of insertions.

Table 3. Ranges for CN-value based on observed/expected reads ratio

Observed/expected reads	CN-value	CNV interpretation
0 - 0.25	0	Homozygous deletion
0.25 – 0.75	1	Heterozygous deletion
0.75 – 1.25	2	Neutral

1.25 – 1.75	3	Heterozygous duplication
1.75 – 2.25	4	Double duplication
> 2.25	5	Other duplications

3. *ACE2* gene variants may underlie interindividual variability and susceptibility to COVID-19 in the Italian population

In this chapter, we present the results of our first exploratory analysis where we evaluated if a predisposing genetic background could contribute to the wide interindividual clinical variability. WES data produced by five Italian centers (Siena, Naples, Turin, Bologna, and Rome) interconnected by the Network of Italian Genomes (NIG) were collected to identify variation encompassing the *ACE2* gene, the notoriously SARS-CoV-2 receptor for host cell entry. Computational chemistry methods were used to estimate how the identified *ACE2* variants modify protein stability and SARS-CoV-2 binding. Also, a higher allelic heterogeneity for *ACE2* in controls compared to cases is shown [10]. This premise led us to extend our research by collecting more SARS-CoV-2 infected patients within the GEN-COVID Multicenter Study and by broadening the spectrum of interest to the whole host genetics, as described in the following chapters.

European Journal of Human Genetics (2020) 28:1602–1614
https://doi.org/10.1038/s41431-020-0691-z

ESHG

ARTICLE

ACE2 gene variants may underlie interindividual variability and susceptibility to COVID-19 in the Italian population

Elisa Benetti[1] · Rossella Tita[2] · Ottavia Spiga[3] · Andrea Ciolfi[4] · Giovanni Birolo[5] · Alessandro Bruselles[6] · Gabriella Doddato[7] · Annarita Giliberti[7] · Caterina Marconi[8] · Francesco Musacchia[9] · Tommaso Pippucci[10] · Annalaura Torella[11] · Alfonso Trezza[3] · Floriana Valentino[7] · Margherita Baldassarri[7] · Alfredo Brusco[5,12] · Rosanna Asselta[13,14] · Mirella Bruttini[2,7] · Simone Furini[1] · Marco Seri[8,10] · Vincenzo Nigro[9,11] · Giuseppe Matullo[5,12] · Marco Tartaglia[4] · Francesca Mari[2,7] · GEN-COVID Multicenter Study · Alessandra Renieri[2,7] · Anna Maria Pinto[2]

Received: 27 March 2020 / Revised: 1 June 2020 / Accepted: 30 June 2020 / Published online: 17 July 2020

Abstract

In December 2019, an initial cluster of interstitial bilateral pneumonia emerged in Wuhan, China. A human-to-human transmission was assumed and a previously unrecognized entity, termed coronavirus disease-19 (COVID-19) due to a novel coronavirus (SARS-CoV-2) was described. The infection has rapidly spread out all over the world and Italy has been the first European country experiencing the endemic wave with unexpected clinical severity in comparison with Asian countries. It has been shown that SARS-CoV-2 utilizes angiotensin converting enzyme 2 (ACE2) as host receptor and host proteases for cell surface binding and internalization. Thus, a predisposing genetic background can give reason for interindividual disease susceptibility and/or severity. Taking advantage of the Network of Italian Genomes (NIG), here we mined whole-exome sequencing data of 6930 Italian control individuals from five different centers looking for *ACE2* variants. A number of variants with a potential impact on protein stability were identified. Among these, three more common missense changes, p.(Asn720Asp), p.(Lys26Arg), and p.(Gly211Arg) were predicted to interfere with protein structure and stabilization. Rare variants likely interfering with the internalization process, namely p.(Leu351Val) and p.(Pro389His), predicted to interfere with SARS-CoV-2 spike protein binding, were also observed. Comparison of *ACE2* WES data between a cohort of 131 patients and 258 controls allowed identifying a statistically significant (P value < 0.029) higher allelic variability in controls compared with patients. These findings suggest that a predisposing genetic background may contribute to the observed interindividual clinical variability associated with COVID-19, allowing an evidence-based risk assessment leading to personalized preventive measures and therapeutic options.

Introduction

In December 2019, a new infectious respiratory disease emerged in Wuhan, Hubei province, China [1–3]. An initial cluster of infections likely due to animal-to-human transmission was rapidly followed by a human-to-human transmission [4]. The disease was recognized to be caused by a novel coronavirus (SARS-CoV-2) and termed coronavirus disease-19 (COVID-19). The infection spread within China and all over the world, and it has been declared as pandemic by the World Health Organization (WHO) on 2nd March 2020. The symptoms of COVID-19 range from fever, dry cough, fatigue, congestion, sore throat, and diarrhea to severe interstitial bilateral pneumonia with a ground-glass image at the CT scan. While recent studies provide evidence of a high number of asymptomatic or paucisymptomatic patients who represent the main reservoir for the infection progression, the severe cases can rapidly evolve towards a respiratory distress syndrome which can be lethal [5]. Although age and comorbidity have been described as the

Members and their affiliations of the GEN-COVID Multicenter study group are listed below Acknowledgements.

These authors contributed equally: Elisa Benetti, Rossella Tita

Supplementary information The online version of this article (https://doi.org/10.1038/s41431-020-0691-z) contains supplementary material, which is available to authorized users.

✉ Alessandra Renieri
alessandra.renieri@unisi.it

Extended author information available on the last page of the article

SPRINGER NATURE

main determinants of disease progression towards severe respiratory distress, the high variation in clinical severity among middle-age adults and children would likely suggest a strong role of the host genetic asset.

A high sequence homology has been shown between SARS-associated coronavirus (SARS-CoV) and SARS-CoV-2 [6]. Recent studies modeled the spike protein to identify the receptor for SARS-CoV-2 and indicated that angiotensin converting enzyme 2 (ACE2) is the receptor for this novel coronavirus [7, 8]. Zhou et al. conducted virus infectivity studies and showed that ACE2 is essential for SARS-CoV-2 to enter HeLa cells [9]. Although the binding strength between SARS-CoV-2 and ACE2 is weaker than that between SARS-CoV and ACE2, it is considered as much high as threshold necessary for virus infection. The spike glycoprotein (S-protein), a trimeric glycoprotein in the virion surface (giving the name of crown -*corona* in latin-), mediates receptor recognition throughout its receptor binding domain (RBD) and membrane fusion [10, 11]. Based on recent reports, SARS-CoV-2 protein binds to ACE2 through Leu455, Phe486, Gln493, Ala501, and Tyr505. It has been postulated that residues 31, 41, 82, 353, 355, and 357 of the ACE2 receptor map to the surface of the protein interacting with SARS-CoV-2 spike protein [12], as previously documented for SARS-CoV. Following interaction, cleavage of the C-terminal segment of ACE2 by proteases, such as transmembrane protease serine 2 (TMPRSS2), enhances the spike protein-driven viral entry [13, 14]. Thus, it is possible, in principle, that genetic variability of the ACE2 receptor is one of the elements modulating virion intake and thus disease severity. *ACE2* is located on chromosome X. Although it is one of the genes escaping X inactivation several lines of evidence suggest that a different degree of X-chromosome inactivation (XCI) is present in distinct tissues [15].

Taking advantage of the Network of Italian Genomes (NIG), a consortium established to generate a public database (NIG-db) containing aggregate variant frequencies data for the Italian population (http://www.nig.cineca.it/), here we describe the genetic variation of *ACE2* in the Italian population, one of the newly affected countries by the SARS-CoV-2 outbreak causing COVID-19. Three common c.2158A>G p.(Asn720Asp), c.77A>G p.(Lys26Arg), and c.631G>A p.(Gly211Arg) variants and 27 rare missense variants were identified, 9 of which had not previously been reported in public databases. We show that p.(Asn720Asp), which lies in a residue located close to the cleavage sequence of TMPRSS2, likely affects the cleavage-dependent virion intake. Along with the other two common variants, this substitution is represented in the Italian and European populations but is extremely rare in the Asian population. We also show that two rare variants, namely, c.1051C>G p.(Leu351Val) and c.1166C>A p.(Pro389His) are predicted to cause conformational changes impacting RBD interaction. As the uncertainty regarding the transmissibility and severity of disease rise, we believe that a deeper characterization of the host genetics and functional characterization of variants may help not only in understanding the pathophysiology of the disease but also in envisaging risk assessment.

Materials and methods

Italian population randomization

The work has been realized in the context of the NIG, with the contribution of centers: Azienda Ospedaliera Universitaria Senese, Azienda Ospedaliera Universitaria Policlinico Sant'Orsola-Malpighi di Bologna, Città della Salute e della Scienza di Torino, Università della Campania "Luigi Vanvitelli", Ospedale Pediatrico Bambino Gesù. The NIG (http://www.nig.cineca.it/) aim is to create a shared database (NIG-db) containing data from nucleic acids sequencing of Italian subjects. This database allows defining an Italian Reference Genome for the identification of genes responsible for genetic diseases or Italian population susceptibility to complex disorders and for the detection of genetic variants responsible for interindividual differences in disease progression ad /or drug response among the Italian population. Individuals coming to our centers were offered to participate to the NIG study and blood withdrawal was performed upon informed consent. Individuals provided signed informed consents at each participating center for whole-exome sequencing analysis (WES), and clinical and molecular data storage and usage. All subjects were unrelated, healthy, and of Italian ancestry. Italian origin was ascertained asking for parents and grandparents origin. DNA has been stored in the Telethon Network of Genetic Biobanks (project no. GTB12001), funded by Telethon Italy.

COVID-19 patients enrollment

The study was consistent with Institutional guidelines and approved by the University Hospital (Azienda Ospedaliera Universitaria Senese) Ethical Committee, Siena, Italy (Prot n. 16929, dated March 16, 2020). Written informed consent was obtained from all patients and controls. Peripheral blood samples in EDTA-containing tubes and detailed clinical data were collected. All these data were inserted in a section of the established and certified Biobank and Registry of the Medical Genetics Unit of the Hospital dedicated to COVID-19. The cohort of COVID-19 patients consists of 131 individuals out of whom 34 females and 97 males belonging to the GEN-COVID MULTICENTER STUDY

([16], Late Breaking Abstract ESHG 2020.2 Virtual Conference "WES profiling of COVID-19"). The cohort of controls consists of 258 italian individuals (129 males and 129 females). All patients are of Italian ethnicity. The median age is 64 years (range 31–98): median age for women 66 years and for males 63 years. The population was clustered into four qualitative severity groups depending on the respiratory impairment and the need for ventilation: high care intensity group (those requiring invasive ventilation), intermediate care intensity group (those requiring noninvasive ventilation i.e., CPAP and BiPAP, and high-flows oxygen therapy), low care intensity group (those requiring conventional oxygen therapy) and very low care intensity group (those not requiring oxygen therapy).

Whole-exome sequencing

Targeted enrichment and massively parallel sequencing were performed on genomic DNA extracted from circulating leukocytes of 6930 individuals. Genomic DNA was extracted from peripheral blood samples using standard procedures. Exome capture was carried out using SureSelect Human All Exon V4/V5/V6/V7 (Agilent Technologies, Santa Clara, CA), Clinical Research Exome V1/V2 (Agilent), Nextera Rapid Capture v.1.2 (Illumina, San Diego, CA), TruSeq Exome Targeted Regions (Illumina, San Diego, CA), TruSight One Expanded V2 (Illumina, San Diego, CA), Sequencing-by-Synthesis Kit v3/v4 (Illumina, San Diego, CA) or HiSeq 2000 v2 Sequencing-by-Synthesis Kit (Illumina, San Diego, CA), and sequencing was performed on Genome Analyzer (v3/v4)/HiSeq2000/NextSeq550/NextSeq500/Novaseq6000 platforms (Illumina, San Diego, CA). A subset of WES had been outsourced (BGI, Shenzhen, China; Mount Sinai, NY, USA; Broad Institute, Harvard, USA). Alignment of raw reads against reference genome Hg19, variant calling and annotation were attained using in-house pipelines [17–19] which take advantage of the GATK Best Practices workflow [20] and of Annovar, VEP [21, 22]. The genome aggregation database gnomAD (https://gnomad.broadinstitute.org/) was used to assess allele frequency for each variant among different populations. The mean depth of coverage of each *ACE2* exon in all participants was 55×. Variants with a depth of coverage lower that 20× were filtered out according to ASHG Guidelines for germline variants [23].

The identified variants have been submitted in LOVD database:

Variant ID 0000667129 https://databases.lovd.nl/shared/individuals/00302622;

Variant ID 0000667137 https://databases.lovd.nl/shared/individuals/00302630;

Variant ID 0000667136 https://databases.lovd.nl/shared/individuals/00302628;

Variant ID 0000667138 https://databases.lovd.nl/shared/individuals/00302629;

Variant ID 0000667131 https://databases.lovd.nl/shared/individuals/00302624;

Variant ID 0000667133 https://databases.lovd.nl/shared/individuals/00302626;

Variant ID 0000667130 https://databases.lovd.nl/shared/individuals/00302621;

Variant ID 0000667134 https://databases.lovd.nl/shared/individuals/00302625;

Variant ID 0000667132 https://databases.lovd.nl/shared/individuals/00302623;

Variant ID 0000667128 https://databases.lovd.nl/shared/individuals/00302620;

Variant ID 0000667126 https://databases.lovd.nl/shared/individuals/00302618;

Variant ID 0000667127 https://databases.lovd.nl/shared/individuals/00302619;

Variant ID 0000667125 https://databases.lovd.nl/shared/individuals/00302617;

Variant ID 0000667123 https://databases.lovd.nl/shared/individuals/00302615;

Variant ID 0000667124 https://databases.lovd.nl/shared/individuals/00302616;

Variant ID 0000667118 https://databases.lovd.nl/shared/individuals/00302610;

Variant ID 0000667120 https://databases.lovd.nl/shared/individuals/00302612;

Variant ID 0000667122 https://databases.lovd.nl/shared/individuals/00302614;

Variant ID 0000667121 https://databases.lovd.nl/shared/individuals/00302613;

Variant ID 0000667119 https://databases.lovd.nl/shared/individuals/00302611;

Variant ID 0000667117 https://databases.lovd.nl/shared/individuals/00302609.

Computational studies

The structure of native human angiotensin converting enzyme-related carboxypeptidase (ACE2) was downloaded from Protein Data Bank (https://www.rcsb.org/) (PDB ID code 1R42) [24]. The DUET program [25] was used to predict the possible effect of amino acids substitutions on the protein structure and function, based on the use of machine-learning algorithms exploiting the three-dimensional structure to quantitatively predict the effects of residue substitutions on protein functionality. Molecular dynamics (MD) simulations of wild-type and variant ACE2 proteins were carried out in GROMACS 2019.3 [26] to calculate root mean square deviation (RMSD) to define structural stability. The graphs were plotted by the

XMGrace software [27]. MD simulations were performed using a high parallel computing infrastructure (HPCS) with 660 cpu within 21 different nodes, 190T of RAM, 30T hard disk partition size, and six NVIDIA TESLA gpu with CUDA support. PyMOL2.3 was used as a molecular graphic interface. The protein structures were solvated in a triclinic box filled with TIP3P water molecules and Na^+/Cl^- ions were added to neutralize the system. The whole systems were then minimized with a maximal force tolerance of 1000 kJ mol^{-1} nm^{-1} using the steepest descendent algorithm. The optimized systems were gradually heated to 310 K in 1 ns in the NVT ensemble, followed by 10 ns equilibration in the NPT ensemble at 1 atm and 310 K, using the V-Rescale thermostat and Berendsen barostat [28, 29]. Subsequently, a further 100 ns MD simulations were performed for data analysis.

Results

ACE2 variants identification

The extent of variability along the entire *ACE2* coding sequence and flanking intronic stretches was assessed using 6930 Italian WES, out of which 4171 males and 2759 females which sum up to 9689 alleles. Identified variants and predicted effects on protein stability are summarized in Tables 1, 2, and Table S1, and represented in Fig. 1. Three more common variants, c.2158A>G p.(Asn720Asp), c.77A>G p.(Lys26Arg), c.631G>A p.(Gly211Arg), were identified. The c.2158A>G p.(Asn720Asp) substitution was estimated to have a frequency of 0.011 (103/9689 alleles), which is in line with the frequency of the variant reported in the gnomAD database (0.016), and is lower than the frequency reported in gnomAD for the European non-Finnish population (0.025, 2195/87966 analyzed alleles). Given the *ACE2* localization on X chromosome we focused our attention on the females alleles. All analyzed females (2759 out of 6930) belonging to the Italian population, were heterozygotes for the variant. Notably, this variant has not been reported in the Eastern Asia population (13,784 exomes). The c.77A>G p.(Lys26Arg), c.631G>A p.(Gly211Arg) variants were found with a frequency of 0.0011 (lower than the frequency in the European non-Finnish population, 0.0058) and 0.0012 (European non-Finnish population frequency, 0.0019), respectively. Out of ~92,708 analyzed alleles in the European non-Finnish population, one homozygous female has been reported for the c.77A>G p.(Lys26Arg) while no homozygous females were reported for the c.631G>A p.(Gly211Arg). According to gnomAD, the allele frequency of the c.77A>G p.(Lys26Arg) variant in the Eastern Asia population was 6×10^{-5}, while the c.631G>A p.(Gly211Arg) has not been reported in 14.822 exomes. In addition to these variants, 28 rare missense variants were identified, out of which ten had not previously been reported in GnomAD database and nine truncating variants that had not been reported in gnomAD database (Table 1 and Supplementary Table 1). Out of these variants, two fall in the neck domain, which is essential for dimerization and one in the intracellular domain. Many of them truncate the protein in different positions of the Protease domain embedded in the extracellular domain, which contains the receptor binding site for SARS-CoV-2. Only three truncating variants have been previously described for *ACE2* likely due to a low-tolerance for loss-of-function variants. In line with this evidence, all these variants were very rare and no homozygous females were detected for the identified variants. Three missense changes c.1517T>C p.(Val506Ala), c.626T>G p.(Val209Gly), and c.1129G>T p.(Gly377Glu) were predicted to have destabilizing structural consequences (Table 2); among these, c.1517T>C p.(Val506Ala) is indeed the only amino acid change reported in the European non-Finnish population (rs775181355; allele frequency 1.40×10^{-5}, CADD 27,2) and is predicted as probably damaging for the protein structure by Polyphen and deleterious by SIFT. Similarly, c.1051C>G p.(Leu351Val) and c.1166C>A p.(Pro389His), which affect a highly hydrophobic core, were predicted to induce conformational changes influencing the interaction with spike protein. The amino acid substitution c.1166C>A p.(Pro389His) (rs762890235, European non-Finnish population allele frequency: 2.45×10^{-5}, CADD 24,8) was predicted to be probably damaging by Polyphen and deleterious by SIFT. Moreover, this rare variant has never been reported in Asian populations.

ACE2 variants likely affect protein stability and SARS-CoV-2 binding

MD analysis provides bona fide simulations of protein structural changes caused by missense variants effects. Yet, its computationally expensive procedure led us to perform MD simulation for only a selection of representative candidate variants. Indeed, we selected the following five variants and corresponding effects: c.1517T>C p.(Val506Ala) which has the higher destabilizing effect, c.77A>G p.(Lys26Arg) and c.631G>A p.(Gly211Arg) with higher allele frequency along with c.1051C>G p.(Leu351-Val) and the c.1166C>A p.(Pro389His) with a predicted effect on spike protein interaction. To analyse differences in protein structure between wild type and mutants, we performed 100 ns MD simulations. The comparison was performed by RMSD analysis. The global effects of the residue substitutions on flexibility and global correlated motion of ACE2 protein are represented in Fig. 2 and the simulation is provided in Supplementary Video S1, S2, S3, S4, S5. While

Table 1 (continued)

NM_021804.2 (hg19)													
Genomic position	Nucleotide change	Amino acid change	CADD_phred	dbSNP	gnomAD	Hemizygous M	Hetenozygous F	Homozygous F	N° of events\n° of alleles	NIG allele frequency	mean DP
X:15613038	c.275C>T	p.(Thr92Ile)	0.031	rs763395248	0.000011	–	2	–	2/9689	0.00020	133
X:15613119	c.194C>T	p.(Ala65Val)	11.7	–	–	–	1	–	1/9689	0.00010	21
X:15618872	c.163A>G	p.(Thr55Ala)	23.8	rs775273812	0.0000057	1	–	–	1/9689	0.00010	214
X:15618958	c.77A>G	p.(Lys26Arg)	10.5	rs4646116	0.0039	4	7	–	11/9689	0.00110	135
X:15619013	c.22C>T	p.(Leu8Phe)	14.2	rs201035388	0.000076	1	1	–	2/9689	0.00020	111

The table reports the genomic position, the nucleotidic, and protein change of exonic *ACE2* identified variants. The genomic reference sequence is NM_021804.2 (hg19). CADD_phred scores are reported for the missense variants. When available, dbSNP rs number and the genome aggregation database gnomAD allele frequency are reported. For all variants are reported the number of individuals hemizygous, heterozygous, or homozygous.

a similar trend for wild-type, c.77A>G p.(Lys26Arg) and c.1517T>C p.(Val506Ala) was observed with a steady course in the RMSD value, which stabilizes at an average of 0.2, 0.25, and 0.3 nm, respectively (Fig. 3a), the c.1166C>A p.(Pro389His) and c.1051C>G p.(Leu351Val) variants show a difference in comparison with the native protein with a gradual increase in RMSD value, which stabilizes at an average of 0.5 nm (Fig. 3a). Finally, the c.631G>A p.(Gly211Arg) shows a bigger difference with a higher increase in RMSD value, which stabilizes at an average of 0.6 nm (Fig. 3a). Structural analysis between WT and mutant c.1517T>C p.(Val506Ala) MD simulations showed that the c.1517T>C p.(Val506Ala) forms a hydrophobic center together with Leu456, Leu503, and Phe516 with minimum differences in protein rearrangements when the residue is mutated in Ala as reported in Fig. 2 and Supplementary Video S1. The c.77A>G p.(Lys26Arg) is located at the N-terminus and the sidechain engages a hydrogen bond with Asn90 thus determining a minimal destabilizing predicted effect as shown in Table 2 and confirmed by RMSD analysis. The c.1166C>A p.(Pro389His) and the c.1051C>G p.(Leu351Val) variants, located in the region for the spike protein interaction, are characterized by a partially overlapping destabilizing effect. The c.1166C>A p.(Pro389His) variant sidechain being more bulky causes the shift of ACE2 (30–40) helix involved in spike protein interaction which being freer to move engages an interaction with Gln96 (Fig. 2 and Supplementary video S5). The c.1051C>G p.(Leu351Val) shorter sidechain is enable to interact with the hydrophobic core composed by Trp349 and Leu45 with a consequent rearrangement of the protein conformation. Finally, while c.631G>A p.(Gly211Arg) has theoretically a smaller destabilizing effect because of an external sidechain which is not involved in particular interaction network, as shown by MD simulation, it confers a wide flexibility to this region because the polar sidechain is able to engage different interactions with vicinal amino acid residues (Fig. 2 and Supplementary Video S2). During MD simulations, we have also investigated the surrounding region of ACE2 WT and previously selected variants by calculating change in solvent accessibility surface area (SASA). Differences in average SASA value would suggest for the native protein a wider surface exposed to solvent and subsequently a different ability to interact with spike SARS-CoV-2 in comparison with the studied variants (Fig. 3b).

Differences in *ACE2* allelic variability in COVID-19 patients compared with controls

In order to shed light on the role of ACE2 variants on interindividual variability and susceptibility to COVID-19 in Italian population we performed WES analysis on a cohort of 131 patients and 258 controls who agreed in

Table 2 Predicted changes in ACE2 protein stability as consequence of residues changes.

Wild Residue	Residue position	Mutant Residue	Predicted ΔΔG	Interaction Network around (5 Å)	Outcome
V	506	A	-2,456	*Y180, L456, R460, P500, A501, S502, L503, F504, H505, N506, S507*	*Highly Destabilizing*
V	209	G	-2,353	*Y207, E208, V209, N210, G211, V212, Y215, D216, Y217, P565, T567*	*Highly Destabilizing*
G	377	E	-2,231	*H373,* H374, E375, M376, G377, H378, I379, A380, Y381, F315, H401, V404, G405, M408	*Highly Destabilizing*
A	264	G	-1,555	*L262, P263, A264, H265, L266, L267, W271, W478, V487, V488, E489, P490, W165*	*Destabilizing*
C	498	R	-1,539	*Y497, C498, D499, P500, A501, S502, G173, R177, L176, Y180, W459, W473, M474*	*Destabilizing*
A	246	T	-1,454	*A242, Y243, V244, R245, A246, K247, L248, M249*	*Destabilizing*
G	377	W	-1,318	*H373 .* H374, E375, M376, G377, H378, I379, A380, Y381, F315, H401, V404, G405, M408	*Destabilizing*
L	351	V	-1,173	*W349, D350, L351, G352 ,* D355, R357, Y41, S44, L45, W48	*Destabilizing*
P	389	H	-1,161	*A387, Q388, P389, F390, L391, L392,* N33, T92, Q96	*Destabilizing*
T	55	A	-0,948	N53, I54, T55, E56, E57, N58, V59	*Destabilizing*
D	206	G	-0,87	*W203, G205, D206, Y207, A396, N397, E398, G399*	*Destabilizing*
K	26	R	-0,79	*E22, E23, Q24, A25, K26, T27, L29, N90, V93*	*Destabilizing*
N	580	D	-0,629	*M579, N580, V581, R582, P583, Q524*	*Destabilizing*
S	547	C	-0,611	*I544, S545 ,* N546, S547, T548, E549, A550, G551	*Destabilizing*
A	65	V	-0,423	*N61, M62, N63, N64, A65, G66, D67, K68, Q42, S43, S44, A46*	*Destabilizing*
H	505	R	-0,345	*L503, F504, M505, F512, Y515, Y510, S511, R273*	*Destabilizing*
T	92	V	-0,322	N90, L91, T92, V93, L95, Q96, P389, L392, S563, E564	*Destabilizing*
E	329	G	-0,302	*Q325, G326, F327, W328, E329, N330, S331*	*Destabilizing*
G	211	R	-0,283	*V209, N210, G211, V212, D213, D216*	*Destabilizing*
T	92	I	-0,155	N90, L91, T92, V93, L95, Q96, P389, L392, S563, E564	*Destabilizing*
D	494	V	-0,041	*H493, D494, E495, T496, Y497*	*Destabilizing*
Q	102	P	0,036	Q98, A99, Q102, N103, G104	(Stabilizing)

DUET program results that display predicted change in folding free energy upon ACE2 missense variant (ΔΔG in kcal/mol). In the first three columns are reported single missense variants with specific position on ACE2 protein. The residues in the first column highlighted in gray are involved in N-glycosylation pattern NxT/S, therefore those missense variants determine the loss glycosylation of Asparagine 53 and 90, respectively. In the fourth column is reported ΔΔG analysis predict effects of missense variants on protein stability using an integrated computational approach. The column "Interaction Network around (5 Å)" shows for each single missense variant the residues around 5 Å. In this column, we highlight in green residues involved in spike SARS-CoV protein interaction, in yellow residues involved in Zinc coordination and finally in magenta residues of Asn involved in N-glycosylation. The last column defines the outcome of protein stability for each single missense variant. An increasing negative value for the ΔΔG is correlated with a higher destabilizing effect, while a positive value is associated with a variant predicted as stabilizing.

participating to the study (see "Materials and methods"). Data analysis of *ACE2* variants identified a different distribution of variants in controls compared with patients (Fig. 4) with the c.2158A>G p.(Asn720Asp) variant being present in two hemizygous male patients (allele frequency 0.012) compared with seven heterozygous female and four hemizygous male controls (allele frequency 0.028). A silent variant the c.2247G>A p.V749V, was also detected in 26 control individuals (allele frequency 0.069) compared with five COVID-19 patients (allele frequency 0.030). Although any single variant was not statistically significantly enriched in one cohort compared to the other, a cumulative analysis of the identified variants detected a statistically significant higher *ACE2* allelic variability (*P* value <0.029) in the control group compared with the patient cohort.

Discussion

According to recent reports, ACE2 is essential for SARS-CoV-2 to enter cells. Recent single-cell RNA studies have also shown that ACE2 is expressed in human lung cells [30]. The majority of ACE2-expressing cells are alveolar type 2 cells. Other ACE2-expressing cells include alveolar type 1 cells, airway epithelial cells, fibroblasts, endothelial cells, and macrophages although their ACE2-expressing cell ratio is low and variable among individuals. The expression and distribution of the ACE2 receptor can thus justify the route of infection and the main localization at the alveolar level. Although the different density of ACE2 receptors in the upper respiratory tract among individuals can partially give reason for the clinical variability, which

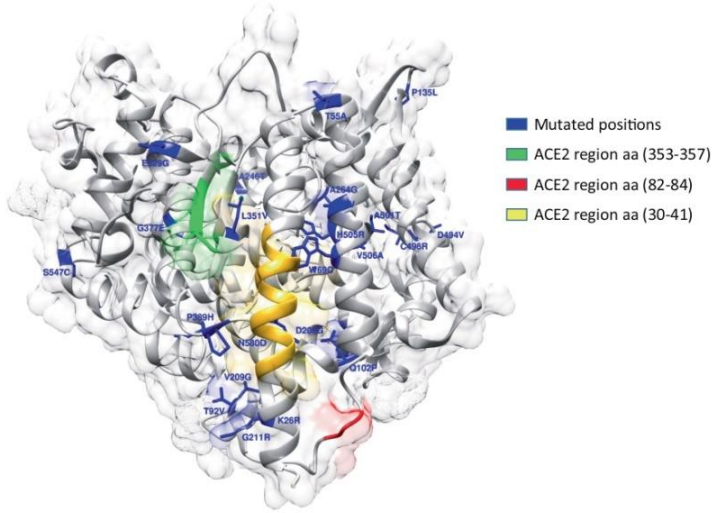


Fig. 1 ACE2 crystal structure with PDB ID 1R42. Surface and cartoon representations of protein in gray. In blue stick are represented each single mutated positions, cartoon region in yellow represent segment between amino acid 30–41, cartoon in green represent segment between amino acid 353–357 and cartoon in red represent segment between amino acid 82–84 that are protein regions responsible of interaction with 2019-nCOv spike glycoprotein.

ranges from asymptomatic/paucisymptomatic patients to severely affected ones, it could not be the only reason for such variability. In addition, recent works did not observe significantly different viral loads in nasal swabs between symptomatic and asymptomatic patients [31]. Italy has been the first European country that experienced the COVID-19 outbreak with a rapid increase in the positive cases in a very short-time period and a morbidity and lethality (~10%) definitely higher in comparison with Asian countries, such as China (4%) and South Korea (1.2%) [32]. These considerations raise the possibility of a predisposing genetic background accounting for or contributing to the wide interindividual clinical variability, as well for the differential morbidity and lethality observed among different countries, population awareness, and constrictive measures apart.

We integrated genomic WES data produced by five Italian centers (Siena, Naples, Turin, Bologna, and Rome) interconnected by the NIG in the attempt to identify variation encompassing the *ACE2* gene, which could account for a difference in SARS-CoV-2 spike binding affinity, processing, or internalization. Previous studies showed that the residues near lysine 31, and tyrosine 41, 82–84, and 353–357 in human ACE2 are important for the binding of S-protein to coronavirus [12]. In line with previous reports [33], we did not find polymorphism or rare variants in these residues in the Italian population. However, we identified three variants namely c.2158A>G p.(Asn720Asp), c.1166C>A p.(Pro389His), and c.1051C>G p.(Leu351Val), one of which polymorphic c.2158A>G p.(Asn720Asp), moderately expressed in the Italian and European non-Finnish populations and with a very low allele frequency or not occurring in the Eastern Asia population. These variants which surround residual essentials for the SARS-CoV-2 spike protein binding were predicted to likely affect the cleavage-dependent virion intake, such as the polymorphic c.2158A>G p.(Asn720Asp) (allele frequency 0.011) which lies four amino acids from the cleavage sequence of TMPRSS2 or to have a substantial impact on protein structure and spike protein interaction by MD simulation (Fig. 3a). The relatively frequent c.631G>A p.(Gly211Arg) (allele frequency 0.0012, 12/6930 individuals) was predicted to confer a wide flexibility to the region because of the ability to engage different interactions with the nearby amino acid residues. Along with these more common variants we also identified very rare variants such as the c.1166C>A p.(Pro389His) and the c.1051C>G p.(Leu351Val), some of which only described in the non-

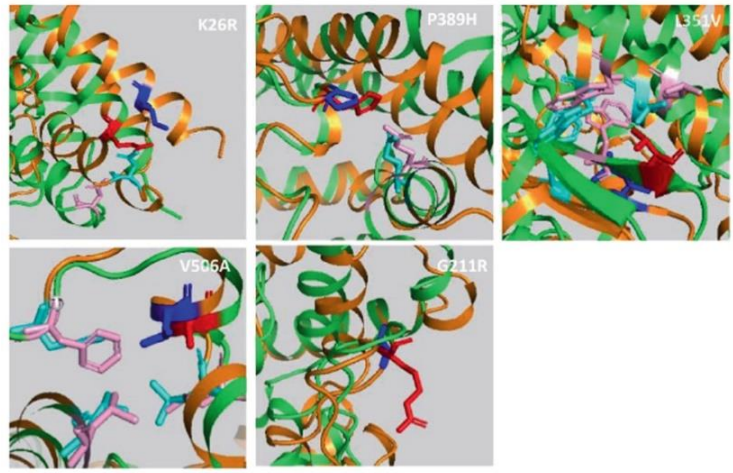


Fig. 2 ACE2 wild-type and variants superimposed structures after 100 ns MD simulation. Cartoon representation of ACE2 wild type (orange) and variants (green) in blue sticks the wild-type residues while in red the corresponding variants. In cyan and pink sticks residues interacting with each specific position.

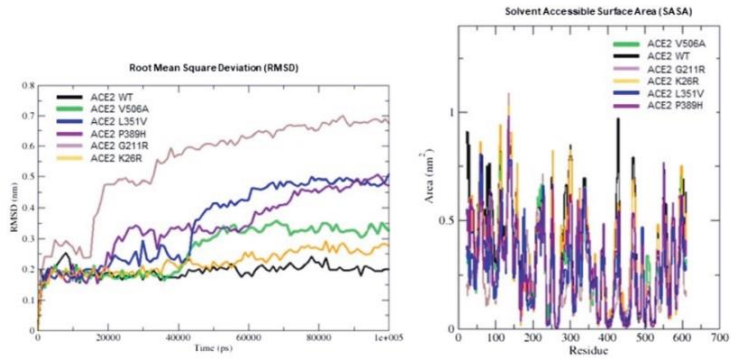


Fig. 3 Structure superimposition snapshot between wild-type protein and variant proteins. a Root mean square deviation (RMSD) trends for the backbone of ACE2 WT (black line) and some selected variants (colored lines, see legend) during 100 ns of simulation. The molecular dynamics simulation shows a good stability for all systems with exception of G211R mutants. RMSD is a parameter used to define the stability of an element. Wild type shows a steady course in the RMSD value, stabilizing at an average of 0.2 nm, while, the G211R variant shows a gradual increase in RMSD value, stabilizing at an average of 0.6 nm. **b** SASA graphical representation of ACE2 WT (black line) and ACE2 variants (colored lines, see legend).

Finnish European population, that could give reason for a different affinity for the SARS-CoV-2 spike protein (Figs. 2, 3a and Supplementary Video S4). Interestingly all the studied variants affect residues highly conserved among species (Supplementary Fig. S1). Given their rarity in other populations, we cannot exclude that these variants can

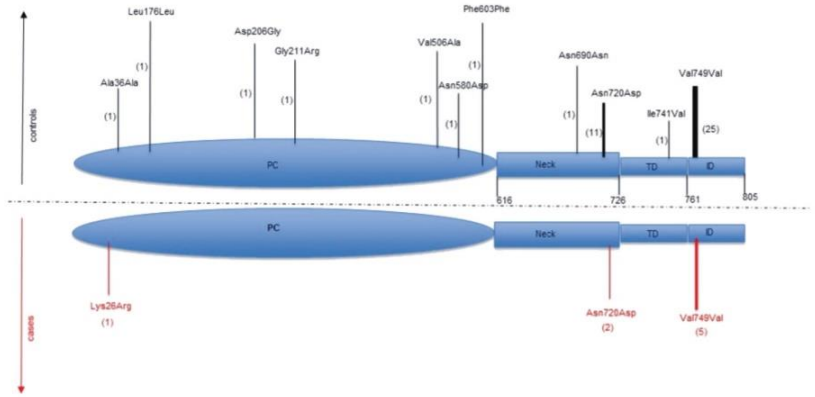


Fig. 4 Differences in ACE2 variants in COVID-19 patients compared with controls. The figure shows the variants located in the ACE2 protein domains. The variants present in controls are shown in black while the variants in cases are shown in red. The number of patients carrying the variant is shown in brackets.

partially account for the clinical outcome observed in the Italian population. WES data generated from a wide cohort of COVID-19 Italian patients revealed a statistically significant ($P < 0{,}029$) higher allelic heterogeneity for *ACE2* in controls compared with patients with a higher chance to find at least one *ACE2* variant in the cohort of controls compared with the cohort of patients. Therefore, it is plausible to think that the effect of allelic variability on ACE2 conformation would at least partially account for the interindividual clinical differences and likely modulate clinical severity. This finding reinforces the hypothesis that at least some of the identified variants or the cumulative effect of few of them confer a different susceptibility to virus cell entry and consequently to disease onset and progression. We cannot exclude that also silent variants such as the c.2247G>A (p. Val749Val) with no effect on the protein could play a role because of an unpredictable impact at a posttranscriptional level.

Notably, morbidity and lethality have been reported definitely higher in men compared with women (~70% vs. 30%, 20th March 2020 ISS report). Although several parameters have been brought to case to explain this difference, i.e., smoking, differences in ACE2 localization and/or density in alveolar cells, hormonal asset, it is noteworthy that *ACE2* is located on chromosome X and that given the low allele frequency of the identified variants the rate of homozygous women is extremely low (see Results section). The XCI is incomplete in humans and some genes show a degree of XCI escape which vary between individuals and tissues [34]. *ACE2* is one of the genes escaping X inactivation, but it belongs to a subgroup of X-chromosome genes showing a higher expression in men in several tissues thus mostly suggesting that ACE2 gene XCI is present although different in distinct tissues [15]. Therefore, the impact of X inactivation on the alternate expression of the two alleles would guarantee, in the affected tissues, a heterogeneous population of ACE2 molecules, some of which protective towards the infection until the point of a complete or almost complete protection in the case of a X inactivation skewed towards the less SARS-CoV-2-binding prone allele. This hypothesis would justify the high rate of asymptomatic or paucisymptomatic patients. However, the presented data does not allow to confirm a clear cause–effect relationship and, since most of the identified variants have very low frequencies, further functional studies are needed to validate these results. *ACE2* is definitely one of the main molecules whose genetic heterogeneity can modulate infection and disease progression; however, a deeper characterization of the host genetics and functional variants in other pathway-related genes may help in understanding the pathophysiology of the disease opening up the way to a stratified risk assessment and to tailored preventive measures and treatments.

Acknowledgements This work was, in part, supported by: Telethon Network of Genetic Biobanks (project no. GTB12001), funded by Telethon Italy; Fondazione Bambino Gesù (*Vite Coraggiose* to MT); Mount Sinai, NY (USA) in the context of the international project ASC (Autism sequencing consortium); SOLVE-RD and MIUR project "Dipartimenti di Eccellenza 2018–2022" (n° D15D18000410001) to the Department of Medical Sciences, University of Torino (GM and

AB). We thank the CINECA consortium for providing computational resources. This study is part of GEN-COVID, https://sites.google.com/dbm.unisi.it/gen-covid, the Italian multicenter study aimed to identify the COVID-19 host genetic bases.

GEN-COVID Multicenter Study Elisa Frullanti[15], Chiara Fallerini[15], Sergio Daga[15], Susanna Croci[15], Sara Amitrano[16], Francesca Fava[15,16], Francesca Montagnani[17,19], Laura Di Sarno[15], Andrea Tommasi[15,16], Maria Palmieri[15], Arianna Emiliozzi[17,18], Massimiliano Fabbiani[18], Barbara Rossetti[18], Giacomo Zanelli[17,18], Laura Bergantini[19], Miriana D'Alessandro[19], Paolo Cameli[19], David Bennet[19], Federico Anedda[20], Simona Marcantonio[20], Sabino Scolletta[20], Federico Franchi[20], Maria Antonietta Mazzei[21], Edoardo Conticini[22], Luca Cantarini[22], Bruno Frediani[22], Danilo Tacconi[23], Marco Feri[24], Raffaele Scala[25], Genni Spargi[26], Marta Corridi[26], Cesira Nencioni[27], Gian Piero Caldarelli[28], Maurizio Spagnesi[29], Paolo Piacentini[29], Maria Bandini[29], Elena Desanctis[29], Anna Canaccini[30], Chiara Spertilli[23], Alice Donati[24], Luca Guidelli[25], Leonardo Croci[27], Agnese Verzuri[30], Valentina Anemoli[30], Agostino Ognibene[31], Massimo Vaghi[32], Antonella D'Arminio Monforte[33], Esther Merlini[33], Mario U. Mondelli[34,35], Stefania Mantovani[34], Serena Ludovisi[34,35], Massimo Girardis[36], Sophie Venturelli[36], Marco Sita[36], Andrea Cossarizza[37], Andrea Antinori[38], Alessandra Vergori[38], Stefano Rusconi[39,40], Matteo Siano[40], Arianna Gabrieli[40], Agostino Riva[39,40], Daniela Francisci[41,42], Elisabetta Schiaroli[41], Pier Giorgio Scotton[43], Francesca Andretta[43], Sandro Panese[44], Renzo Scaggiante[45], Saverio Giuseppe Parisi[46], Francesco Castelli[47], Maria Eugenia Quiros-Roldan[47], Paola Magro[47], Cristina Minardi[47], Deborah Castelli[47], Itala Polesini[47], Matteo Della Monica[48], Carmelo Piscopo[48], Mario Capasso[49,50,51], Roberta Russo[49,50], Immacolata Andolfo[49,50], Achille Iolascon[49,50], Massimo Carella[52], Marco Castori[52], Giuseppe Merla[52], Filippo Aucella[53], Pamela Raggi[54], Carmen Marciano[54], Rita Perna[54], Matteo Bassetti[55,56], Antonio Di Biagio[56], Maurizio Sanguinetti[57,58], Luca Masucci[57,58], Chiara Gabbi[59], Serafina Valente[60], Susanna Guerrini[22], Ilaria Meloni[15], Maria Antonietta Mencarelli[16], Caterina Lo Rizzo[16], Elena Bargagli[20], Marco Mandalà[61], Alessia Giorli[61], Lorenzo Salerni[61], Giuseppe Fiorentino[62], Patrizia Zucchi[63], Pierpaolo Parravicini[63], Elisabetta Menatti[64], Stefano Baratti[65], Tullio Trotta[66], Ferdinando Giannattasio[66], Gabriella Coiro[66], Fabio Lena[67], Domenico A. Coviello[68], Cristina Mussini[69]

[15]Medical Genetics, University of Siena, Siena, Italy; [16]Genetica Medica, Azienda Ospedaliera Universitaria Senese, Senese, Italy; [17]Dept of Medical Biotechnologies, University of Siena, Siena, Italy; [18]Dept of Specialized and Internal Medicine, Tropical and Infectious Diseases Unit, Siena, Italy; [19]Unit of Respiratory Diseases and Lung Transplantation, Department of Internal and Specialist Medicine, University of Siena, Siena, Italy; [20]Dept of Emergency and Urgency, Medicine, Surgery and Neurosciences, Unit of Intensive Care Medicine, Siena University Hospital, Siena, Italy; [21]Department of Medical, Surgical and Neuro Sciences and Radiological Sciences, Unit of Diagnostic Imaging, University, Siena, Italy; [22]Rheumatology Unit, Department of Medicine, Surgery and Neurosciences, University of Siena, Policlinico Le Scotte, Siena, Italy; [23]Department of Specialized and Internal Medicine, Infectious Diseases Unit, San Donato Hospital Arezzo, Arezzo, Italy; [24]Dept of Emergency, Anesthesia Unit, San Donato Hospital, Arezzo, Italy; [25]Department of Specialized and Internal Medicine, Pneumology Unit and UTIP, San Donato Hospital, Arezzo, Italy; [26]Department of Emergency, Anesthesia Unit, Misericordia Hospital, Grosseto, Italy; [27]Department of Specialized and Internal Medicine, Infectious Diseases Unit, Misericordia Hospital, Grosseto, Italy; [28]Clinical Chemical Analysis Laboratory, Misericordia Hospital, Grosseto, Italy; [29]Department of Prevention, Azienda USL Toscana Sud Est, Arezzo, Italy; [30]Territorial Scientific Technician Department, Azienda USL Toscana Sud Est, Arezzo, Italy; [31]Clinical Chemical Analysis Laboratory, San Donato Hospital, Arezzo, Italy; [32]Chirurgia Vascolare, Ospedale Maggiore di Crema, Crema, Italy; [33]Department of Health Sciences, Clinic of Infectious Diseases, ASST Santi Paolo e Carlo, University of Milan, Milan, Italy; [34]Division of Infectious Diseases and Immunology, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy; [35]Department of Internal Medicine and Therapeutics, University of Pavia, Pavia, Italy; [36]Department of Anesthesia and Intensive Care, University of Modena and Reggio Emilia, Modena, Italy; [37]Department of Medical and Surgical Sciences for Children and Adults, University of Modena and Reggio Emilia, Modena, Italy; [38]HIV/AIDS Department, National Institute for Infectious Diseases, IRCCS, Lazzaro Spallanzani, Rome, Italy; [39]III Infectious Diseases Unit, ASST-FBF-Sacco, Milan, Italy; [40]Department of Biomedical and Clinical Sciences Luigi Sacco, University of Milan, Milan, Italy; [41]Infectious Diseases Clinic, Department of Medicine 2, Azienda Ospedaliera di Perugia and University of Perugia, Santa Maria Hospital, Perugia, Italy; [42]Infectious Diseases Clinic, 'Santa Maria' Hospital, University of Perugia, Perugia, Italy; [43]Department of Infectious Diseases, Treviso Hospital, Local Health Unit 2 Marca Trevigiana, Treviso, Italy; [44]Infectious Diseases Department, Ospedale Civile 'SS. Giovanni e Paolo', Venice, Italy; [45]Infectious Diseases Clinic, ULSS1 Belluno, Italy; [46]Department of Molecular Medicine, University of Padova, Padova, Italy; [47]Department of Infectious and Tropical Diseases, University of Brescia and ASST Spedali Civili Hospital, Brescia, Italy; [48]Medical Genetics and Laboratory of Medical Genetics Unit, A.O.R.N. 'Antonio Cardarelli', Naples, Italy; [49]Department of Molecular Medicine and Medical Biotechnology, University of Naples Federico II, Naples, Italy; [50]CEINGE Biotecnologie Avanzate, Naples, Italy; [51]IRCCS SDN, Naples, Italy; [52]Division of Medical Genetics, Fondazione IRCCS Casa Sollievo della Sofferenza Hospital, San Giovanni Rotondo, Italy; [53]Department of Medical Sciences, Fondazione IRCCS Casa Sollievo della Sofferenza Hospital, San Giovanni Rotondo, Italy; [54]Clinical Trial Office, Fondazione IRCCS Casa Sollievo della Sofferenza Hospital, San Giovanni Rotondo, Italy; [55]Department of Health Sciences, University of Genova, Genova, Italy; [56]Infectious Diseases Clinic, Policlinico San Martino Hospital, IRCCS for Cancer Research Genova, Genova, Italy; [57]Microbiology, Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Catholic University of Medicine, Rome, Italy; [58]Department of Laboratory Sciences and Infectious Diseases, Fondazione Policlinico Universitario A. Gemelli IRCCS, Rome, Italy; [59]Independent Scientist, Milan, Italy; [60]Department of Cardiovascular Diseases, University of Siena, Siena, Italy; [61]Otolaryngology Unit, University of Siena, Siena, Italy; [62]AORN dei Colli Presidio Ospedaliero Cotugno, Azienda, Italy; [63]Department of Internal Medicine, ASST Valtellina e Alto Lario, Sondrio, Italy; [64]Study Coordinator Oncologia Medica e Ufficio Flussi Sondrio, Sondrio, Italy; [65]Department of Infectious and Tropical Diseases, University of Padova, Padova, Italy; [66]First Aid Department, Luigi Curto Hospital, Polla, Salerno, Italy; [67]Local Health Unit-Pharmaceutical Department of Grosseto, Toscana Sud Est Local Health Unit, Grosseto, Italy; [68]U.O.C. Laboratorio di Genetica Umana, IRCCS Istituto G. Gaslini, Genoa, Italy; [69]Infectious Diseases Clinics, University of Modena and Reggio Emilia, Modena, Italy

Author contributions EB, RT, OS, AMP, and AR have made substantial contributions to conception and design, acquisition of data, analysis and interpretation of data, and have been involved in drafting the paper. RA, GB, ABruselles, ABrusco, GD, AG, FM, TP, ATorella, ATrezza, and FV has made substantial contributions to acquisition and analysis of the data. MBaldassarri, MBruttini, AC, SF, FM, GM, VN, MS, and MT have made substantial contributions to interpretation of data and clinical evaluation. All authors have been involved in drafting the paper; have given final approval of the version to be published and agree to be accountable for all aspects of the work in ensuring that

23. Rehm HL, Bale SJ, Bayrak-Toydemir P, Berg JS, Brown KK, Deignan JL, et al. ACMG clinical laboratory standards for next-generation sequencing. Genet Med. 2013;15:733–47. https://doi.org/10.1038/gim.2013.92
24. Towler P, Staker B, Prasad SG, Menon S, Tang J, Parsons T, et al. ACE2 X-ray structures reveal a large hinge-bending motion important for inhibitor binding and catalysis. J Biol Chem. 2004;279:17996–8007. https://doi.org/10.1074/jbc.M311191200
25. Pires D, Ascher DB, Blundell TL. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Res. 2014;42(Web Server issue):W314–W319. https://doi.org/10.1093/nar/gku411
26. Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, et al. Gromacs: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX. 2015;15:19–25. https://doi.org/10.1016/j.softx.2015.06.001
27. Turner PJ XM Grace, Version 5.1.19. Center for Coastal and Land-Margin Research, Oregon Graduate Institute of Science and Technology, Beaverton, OR; 2005.
28. Bussi G, Donadio D, Parrinello M. Canonical sampling through velocity rescaling. J Chem Phys. 2007;126:014101. https://doi.org/10.1063/1.2408420
29. Berendsen HJC, Postma JPM, Van Gunsteren WF, Dinola A, Haak JR. Molecular dynamics with coupling to an external bath. J. Chem. Phys. 1984;81:3684. https://doi.org/10.1063/1.448118
30. Zhao Y, Zhao Z, Wang Y, Zhou Y, Ma Y, Zuo W. Single-cell RNA expression profiling of ACE2, the putative receptor of Wuhan 2019-nCov. bioRxiv. 2020. https://doi.org/10.1101/2020.01.26.919985
31. Cereda D, Tirani M, Rovida F, Demicheli V, Ajelli M, Poletti P, et al. The early phase of the COVID-19 outbreak in Lombardy, Italy. 2020. http://arxiv.org/abs/2003.09320
32. Modi C, Boehm V, Ferraro S, Stein G, Seljak U How deadly is COVID-19? A rigorous analysis of excess mortality and age-dependent fatality rates in Italy. 2020. https://www.medrxiv.org/content/ https://doi.org/10.1101/2020.04.15.20067074v3
33. Cao Y, Li L, Feng Z, Wan S, Huang P, Sun X, et al. Comparative genetic analysis of the novel coronavirus (2019-nCoV/SARS-CoV-2) receptor ACE2 in different populations. Cell Discov. 2020;6:11. https://doi.org/10.1038/s41421-020-0147-1. Published 2020 Feb 24
34. Carrel L, Willard HF. X-inactivation profile reveals extensive variability in X-linked gene expression in females. Nature. 2005;434:400–4. https://doi.org/10.1038/nature03479

Affiliations

Elisa Benetti[1] · Rossella Tita[2] · Ottavia Spiga[3] · Andrea Ciolfi[4] · Giovanni Birolo[5] · Alessandro Bruselles[6] · Gabriella Doddato[7] · Annarita Giliberti[7] · Caterina Marconi[8] · Francesco Musacchia[9] · Tommaso Pippucci[10] · Annalaura Torella[11] · Alfonso Trezza[3] · Floriana Valentino[7] · Margherita Baldassarri[7] · Alfredo Brusco[5,12] · Rosanna Asselta[13,14] · Mirella Bruttini[2,7] · Simone Furini[1] · Marco Seri[8,10] · Vincenzo Nigro[9,11] · Giuseppe Matullo[5,12] · Marco Tartaglia[4] · Francesca Mari[2,7] · GEN-COVID Multicenter Study · Alessandra Renieri[2,7] · Anna Maria Pinto[2]

1 Department of Medical Biotechnologies, University of Siena, Siena, Italy

2 Genetica Medica, Azienda Ospedaliera Universitaria Senese, Siena, Italy

3 Department of Biotechnology, Chemistry and Pharmacy, University of Siena, Siena, Italy

4 Genetics and Rare Diseases Research Division, Ospedale Pediatrico Bambino Gesù, IRCCS, Rome, Italy

5 Department of Medical Sciences, University of Turin, Turin, Italy

6 Department of Oncology and Molecular Medicine, Istituto Superiore di Sanità, Rome, Italy

7 Medical Genetics, University of Siena, Siena, Italy

8 Department of Medical and Surgical Sciences, University of Bologna, Bologna, Italy

9 Telethon Institute of Genetics and Medicine, Pozzuoli, Italy

10 Sant'Orsola-Malpighi University Hospital, Bologna, Italy

11 Dipartimento di Medicina di Precisione, Università della Campania "Luigi Vanvitelli", Napoli, Italy

12 Genetica Medica, Città della Salute e della Scienza, Torino, Italy

13 Department of Biomedical Sciences, Humanitas University, Rozzano, Milan, Italy

14 Humanitas Clinical and Research Center—IRCCS, Rozzano, Milan, Italy

4. Clinical and molecular characterization of COVID-19 hospitalized patients

In the present chapter we provide a comprehensive characterization of COVID-19 hospitalized patients from a clinical and molecular point of view. A multiple-organ involvement is shown, confirming that COVID-19 is a systemic disease rather than just a lung disorder. Considering the great variability of clinical symptoms, the need for a model that could account for both common and rare variants is delineated [50]. The analysis of the contribution of common and rare variants on COVID-19 severity asks for automated procedure that could extract relevant information from the massive datasets derived from WES experiments. Chapters 5, 6, 7 and 8 will provide applications of ML models to address this issue.

PLOS ONE

RESEARCH ARTICLE

Clinical and molecular characterization of COVID-19 hospitalized patients

Elisa Benetti[1], Annarita Giliberti[2], Arianna Emiliozzi[1,3], Floriana Valentino[2], Laura Bergantini[4], Chiara Fallerini[2], Federico Anedda[5], Sara Amitrano[6], Edoardo Conticini[7], Rossella Tita[6], Miriana d'Alessandro[4], Francesca Fava[2,6], Simona Marcantonio[5], Margherita Baldassarri[2], Mirella Bruttini[2,6], Maria Antonietta Mazzei[8], Francesca Montagnani[1,3], Marco Mandalà[9], Elena Bargagli[4], Simone Furini[1], GEN-COVID Multicenter Study¶, Alessandra Renieri[2,6]*, Francesca Mari[2,6]

1 Department of Medical Biotechnologies, University of Siena, Siena, Italy, **2** Medical Genetics, University of Siena, Siena, Italy, **3** Department of Specialized and Internal Medicine, Tropical and Infectious Diseases Unit, Azienda Ospedaliera Universitaria Senese, Senese, Italy, **4** Unit of Respiratory Diseases and Lung Transplantation, Department of Internal and Specialist Medicine, University of Siena, Siena, Italy, **5** Department of Emergency and Urgency, Medicine, Surgery and Neurosciences, Unit of Intensive Care Medicine, Siena University Hospital, Siena, Italy, **6** Genetica Medica, Azienda Ospedaliera Universitaria Senese, Senese, Italy, **7** Rheumatology Unit, Department of Medicine, Surgery and Neurosciences, University of Siena, Policlinico Le Scotte, Siena, Italy, **8** Department of Medical, Surgical and Neuro Sciences and Radiological Sciences, Unit of Diagnostic Imaging, University of Siena, Azienda Ospedaliera Universitaria Senese, Senese, Italy, **9** Otolaryngology Unit, University of Siena, Siena, Italy

¶ Membership of the Termite Genome Working Group is listed in the Acknowledgments.
* alessandra.renieri@unisi.it

OPEN ACCESS

Citation: Benetti E, Giliberti A, Emiliozzi A, Valentino F, Bergantini L, Fallerini C, et al. (2020) Clinical and molecular characterization of COVID-19 hospitalized patients. PLoS ONE 15(11): e0242534. https://doi.org/10.1371/journal.pone.0242534

Editor: Giordano Madeddu, University of Sassari, ITALY

Received: August 7, 2020

Accepted: November 5, 2020

Published: November 18, 2020

Data Availability Statement: Data about the gene-based analyses and variants are available as Supplementary Material. In addition sequencing data are available in the Network for Italian Genomes database (http://www.nig.cineca.it and http://nigdb.cineca.it/).

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Clinical and molecular characterization by Whole Exome Sequencing (WES) is reported in 35 COVID-19 patients attending the University Hospital in Siena, Italy, from April 7 to May 7, 2020. Eighty percent of patients required respiratory assistance, half of them being on mechanical ventilation. Fiftyone percent had hepatic involvement and hyposmia was ascertained in 3 patients. Searching for common genes by collapsing methods against 150 WES of controls of the Italian population failed to give straightforward statistically significant results with the exception of two genes. This result is not unexpected since we are facing the most challenging common disorder triggered by environmental factors with a strong underlying heritability (50%). The lesson learned from Autism-Spectrum-Disorders prompted us to re-analyse the cohort treating each patient as an independent case, following a Mendelian-like model. We identified for each patient an average of 2.5 pathogenic mutations involved in virus infection susceptibility and pinpointing to one or more rare disorder(s). To our knowledge, this is the first report on WES and COVID-19. Our results suggest a combined model for COVID-19 susceptibility with a number of common susceptibility genes which represent the favorite background in which additional host private mutations may determine disease progression.

Introduction

Italy has been the first European Country experiencing the epidemic wave of SARS-CoV-2 infection, with an apparently more severe clinical picture, compared to other countries. Indeed, the case fatality rate has peaked to 14% in Italy, while it remains stable around 5% in China. At the time of the study, 12 May 2020, SARS-CoV-2 positive subjects in Italy have reached the threshold of 200.000 cases [1]. Since the beginning of the epidemic wave, one of the first observations has been a highly heterogeneous phenotypic response to SARS-CoV-2 infection among individuals. Indeed, while most affected subjects show mild symptoms, a subset of patients develops severe pneumonia requiring mechanical ventilation with a 20% of cases requiring hospitalization; 5% of cases admitted to the Intensive Care Unit (ICU), and 6,1% requiring intensive support with ventilators or extracorporeal oxygenation (ECMO) machines [2]. Although patients undergoing ventilatory assistance are often older and are affected by other diseases, like diabetes [3], the existing comorbidities alone do not fully explain the differences in clinical severity. As demonstrated for other viral diseases, the basis of these different outcomes there are host predisposing genetic factors leading to different immunogenicity/cytokine responses as well as specific receptor permissiveness to virus and antiviral defence [4–6]. Similarly, during the study of host genetics in influenza disease, a pattern of genetic markers has been identified which underlies increased susceptibility to a more severe clinical outcome (as reviewed in [7]). This hypothesis is also supported by a recent work reporting 50% heritability of COVID-19 symptoms [8].

The identification of host genetic variants associated with disease severity is of utmost importance to develop both effective treatments, based on a personalized approach, and novel diagnostics. Also, it is expected to be of high relevance in providing guidance for the health care systems and societal organizations. However, nowadays, little is known about the impact of host genome variability on COVID-19 susceptibility and severity.

On March 16th, 2020 the University Hospital in Siena launched a study named GEN-COVID with the aim to collect the genomic DNA of 2,000 COVID-19 patients for host genetic analysis. More than 30 different hospitals and community centers throughout Italy joined the study and are providing samples and clinical detailed information of COVID-19 patients. This study is aimed to identify common and rare genetic variants of SARS-CoV-2 infected individuals, using a whole exome sequencing (WES) analysis approach, in order to establish an association between host genetic variants and COVID-19 severity and prognosis.

Results

Clinical data

The cohort consists of 35 COVID-19 patients (33 unrelated and 2 sisters) admitted to the University Hospital in Siena, Italy, from April 7 to May 7, 2020. All patients are of Caucasian ethnicity, except for one North African and one Hispanic. The mean and median age is 64 years (range 31–98): 11 females (median age 66 years) and 24 males (median age 62 years).

The population is clustered into four qualitative severity groups depending on the respiratory impairment and the need for ventilation (groups 1–4 in Table 1 and different colors in Fig 1) (see Methods section). In the two most severe groups (groups 1 and 2, including 13 patients) there are 11 males and 2 females, while in the two mildest groups (groups 3 and 4 including 22 patients) males are 13 while females are 9.

Patients were also assigned a lung imaging grading according to X-Rays and CT scans. The mean value is 13 for high care intensity group, 12 for intermediate care intensity group, 8 for low care intensity group and 5 for very low care intensity group.

Table 1. Clinical characteristics COVID19 patients admitted to the University Hospital of Siena (Italy).

Subject characteristics	Group 1	Group 2	Group 3	Group 4
No. of subjects (%)	6(17.1%)	7(20%)	15(42.9%)	7(20%)
Mean age (SD)	63 (6.2)	61.6 (12.3)	70 (14)	54 (15.7)
Gender				
Male [n (%)]	5(14%)	6(17%)	7 (20%)	6 (17.1%)
Female[n (%)]	1(2.8%)	1(2.8%)	8 (22.8%)	1(2.8%)
PaO_2/FiO_2 [median (IQR)]				
	94.5 (37.7)	156 (74)	279.5 (162)	304 (73.5)
Lung imaging grading (CXR score)				
[median (IQR)]	13 (3.7)	13 (3)	8 (4)	5(6)
Laboratory findings				
$CD4^+$ T cells count				
[median (IQR)]	300 (330.7)	582 (661)	458 (906)	623 (360)
NK cells count				
[median (IQR)]	79.5 (72.2)	73 (110)	112 (90)	204 (174)
IL-6 value				
[median (IQR)]	*598 (777.7)*	567 (648.2)	14.9 (28.4)	19 (5.3)
Fibrinogen				
[median (IQR)]	*406 (409.7)*	518 (296)	566 (209)	546 (239)
CRP				
[median (IQR)]	*1.22 (24.54)*	0.43 (4.6)	0.36 (1.52)	3.14 (4.97)
LDH				
[median (IQR)]	377 (217)	407 (319)	272 (121)	255 (81)
D-Dimer				
[median (IQR)]	5069.5 (20183)	1526 (54221)	1167 (2022)	884.5 (786.3)
Hyposmia (VAS score) [n (%)]				
<2 (normal)	4(11.3%)	6 (17.1%)	14 (40%)	7(20%)
2–5 (intermediate)	1(2.8%)	0	0	0
>5 (severe)	0	1(2.8%)	1(2.8%)	0
Hypogeusia (VAS score) [n (%)]				
No	4(11.3%)	6(17.1%)	13(37.1%)	7(20%)
Yes	1(2.8%)	1(2.8%)	2(5.7%)	0
Heart involvement [n (%)]				
Yes	4(11.3%)	3(8.6%)	6(17.1%)	0
T = T-Troponin >15 (ng/L); B = pro-BNP M > 88 (pg/ml); F > 153 (pg/ml); A = arrhythmia)	T/B 2(5.7%)	T/B 1(2.8%)	T/B 2(5.7%)	
	B 2(5.7%)	T 1(2.8%)	T/A 1(2.8%)	
		A 1(2.8%)	B/A 1(2.8%)	
			A 1(2.8%)	
			B 1(2.8%)	
No	2(5.7%)	4(11.3%)	9(25.7%)	7(20%)
Unknown	0	0	0	0
Hepatic (H)/Pancreatic involvement (P) [n (%)]				
H and P	2(5.7%)	5(14.3%)	6(17.1%)	1(2.8%)
H only	3(8.6%)	0	1(2.8%)	1(2.8%)
P only	0	0	1(2.8%)	1(2.8%)
None	1(2.8%)	2(5.7%)	7(20%)	4(11.3%)
Kidney involvement [n (%)]				
Yes	0	3(8.6%)	5(14.3%)	1(2.8%)

(*Continued*)

Table 1. (Continued)

Subject characteristics	Group 1	Group 2	Group 3	Group 4
No	6(17.1%)	4(11.3%)	10 (28.6%)	6(17.1%)
Co-morbidities [n (%)]				
Cardiovascular disease	1(2.8%)	2(5.7%)	3(8.6%)	
Hypertension	2(5.7%)	2(5.7%)	8(22.8%)	
Tumor	2(5.7%)	1(2.8%)	2(5.7%)	1(2.8%)
Diabetes			4 (11.3%)	
Pulmonary disease			1(2.8%)	1(2.8%)

COVID-19 cohort is grouped in 4 qualitative severity groups depending on the respiratory impairment and the need of ventilation. Group 1 requires invasive ventilation. Group 2 requires CPAP/BiPAP/high-flows oxygen therapy. Group 3 requires conventional oxygen therapy. Group 4 does not require oxygen therapy. Clinical characteristics are listed and the number of patients are indicated for each of them.

https://doi.org/10.1371/journal.pone.0242534.t001

Regarding immunological findings, a decrease in the total number of peripheral $CD4^+$ T cells were identified in 13 subjects, while NK cells' count was impaired in 10 patients. Six patients showed a reduction of both parameters. IL-6 serum level was elevated in 13 patients.

Hyposmia was present in 3 out of 34 evaluated cases (8.8%), and hypogeusia was present in the same subjects plus another case. These four cases belong to the first three severity groups. Liver involvement was present in 7 cases (20%), while pancreas involvement in 4 cases (11%); 10 patients presented both (29%). Heart involvement was detected in 13 cases (37%). 9 patients (25%) showed kidney involvement. Fibrinogen values below 200 mg/dL were identified in 2 cases (6%), between 200 and 400 mg/dL in 7 cases (20%), and above 400mg/dL in 22 cases (63%). D-dimer value below 500 ng/mL was present in 1 case (3%), between 500 and 5000 ng/mL in 26 cases (74%), and in 7 cases (20%) was 10 times higher than the normal value (>5000 ng/mL) (Table 1).

Unbiased collapsing gene analysis

At first, we tested the hypothesis that susceptibility could be due to one or more common factor(s) in the cohort of patients compared to controls. According to this idea, damaging variants of that/those gene(s) should be either over- or under- represented in patients vs controls. We used, as controls, individuals of the Italian population assuming that the majority of them, if infected, would have shown no severe symptoms. WES data of 35 patients were compared with those of 150 controls (the Siena cohort of the Network of Italian Genomes NIG: http://www.nig.cineca.it) using a gene burden test which compares the rate of disrupting mutations per gene. The variants were collapsed on a gene-by-gene basis, in order to identify genes with mutational burden statistically different between COVID-19 samples and controls. The analysis identified genes harboring deleterious mutations (according to the DANN score) with a statistically significant higher frequency in controls than in COVID-19 patients such as the olfactory receptor gene *OR4C5* (adjusted p-value of 1.5E-10), (Fig 2 and S1 Table) and *NDUFAF7*, although to a lesser extent (Fig 2 and S1 Table). For all these genes, the susceptibility factor is represented by the functioning (or more functioning) gene. We also identified two additional genes, *PRKRA* and *LAPTM4B*, for which the probability of observing a deleterious variant was computed higher in the COVID-19 samples compared to controls (Fig 2 and S2 Table). In these latter cases, the functioning gene represents indeed a protective factor.

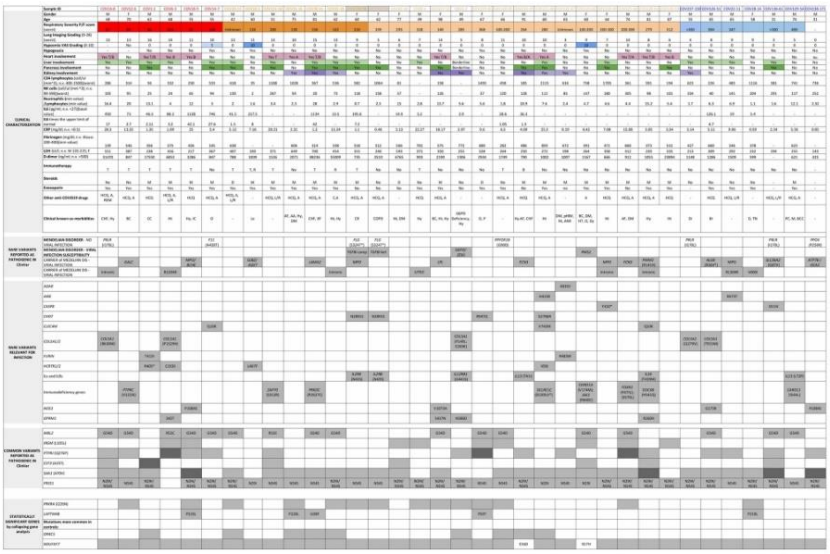


Fig 1. Clinical characteristic and mutated genes. The population is clustered into four qualitative severity groups indicated with different colors depending on the respiratory impairment and the need for ventilation. Red color is used for high care intensity group (those requiring invasive ventilation), orange for intermediate care intensity group (those requiring non invasive ventilation i.e. CPAP and BiPAP, and high-flows oxygen therapy), pink for low care intensity groups (those requiring conventional oxygen therapy) and light blue for very low care intensity groups (those not requiring oxygen therapy). Patients COV132-57 and COV133-58 are reported in grey because they are siblings. A detailed clinical characterization is provided (i.e multiple organs involvement, presence of comorbidity, clinical laboratory parameters, etc.) along with the genetic background for each patient. Liver and pancreas involvement is indicated in green color, heart involvement in magenta and kidney involvement in purple. The presence of hyposmia, classified from 0 to 10 using VAS score, is indicated in light blue. A darker shade of color represents a more severe organ damage. In-silico predicted deleterious variants in genes relevant for infection and pathogenic variants (both common and rare) reported in ClinVar Database are described and a further subdivision between genes involved in a mendelian disorder and/or viral infection susceptibility is provided. For all these gene categories, dark grey is used to identify the homozygous status of the variants while light grey for the heterozygous status. In the end, statistically significant genes obtained after Gene Burden analysis are listed: *PRKRA* and *LAPTM4B* mutational burden resulted to be enriched in the 35 COVID-19 patients compared to the controls, while *OR4C5* and *NDUFAF7* have proven to have an opposite trend, having a mutational burden more enriched in controls. For this reason, for *NDUFAF7*, *OR4C5*, genes the grey color and the white color are inverted because having less variants and by consequence a more functional gene represents a susceptibility factor. For this category, white color underlies a higher mutational burden while grey color indicates lower mutational burden. Clinical known co-morbidities: CHF (Congestive Heart Failure), Hy (Hypertension), CC (Colon Cancer), Le (Leukemia), DM (Diabetes mellitus), pHBV (Previous HBV infection), AMI (Acute myocardial infarction), BC (Breast Cancer), Ht (Hypothyroidism), IC (Ictus Cerebri), O (Obesity), Dy (Dyslipidemia), KF (Kidney Failure), CR (Congenital Rickets), AF (Atrial Fibrillation), Di (Diverticolosis), Br (Bronchiectasis), COPD (Chronic obstructive pulmonary disease), D (Depression), Tn (Thyroid Nodules), PC (Prostate Cancer), M (Melanoma), BCC (Basal cell carcinoma). Immunotherapy: T (Tocilizumab), R (Ruxolitinib), B (Baricitinib). Steroids: M (Methylprednisolone), D (Dexamethasone). Other anti-COVID19 drugs: HCQ (Hydroxicloroquine), A (Azithromycine), REM (Remdesevir), L/R (Lopinavir/Ritonavir), C (Cloroquine).

https://doi.org/10.1371/journal.pone.0242534.g001

Gene analysis using the Mendelian-like model

We then tested the hypothesis that COVID-19 susceptibility is due to different variants in different individuals. A recently acquired knowledge on the genetic bases of Autism Spectrum Disorders suggests that a common disorder could be the sum of many different rare disorders and this genetic landscape can appear indistinguishable at the clinical level [9]. Therefore, we

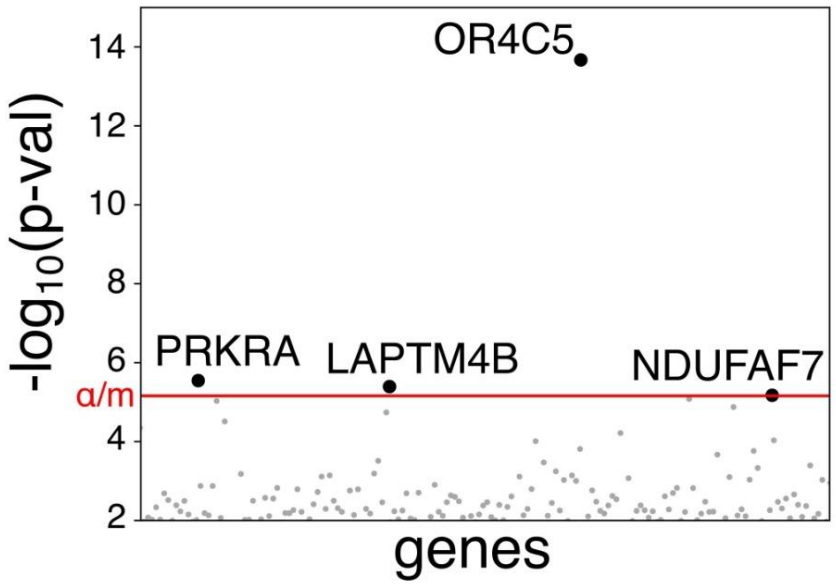


Fig 2. Mutational burden at the gene level. The significance threshold including the Bonferroni correction is shown as a red line ($\alpha = 0.05$, number of test, m = 7196). P-values for genes below the significance threshold are shown as grey dots. Black circles are used for the 4 genes that were identified as statistically different between COVID-19 samples and controls.

https://doi.org/10.1371/journal.pone.0242534.g002

analyzed our cohort treating each patient as an independent case, following a Mendelian-like model. According to the "pathogenic" definition in ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/), for each patient, we identified an average of 1 mutated gene involved in viral infection susceptibility and pinpointing to one or more rare disorder(s) or a carrier status of rare disorders (Fig 1). Following the pipeline used in routine clinical practice for WES analysis in rare disorders we then moved forward checking for rare variants "predicted" to be relevant for infection by the means of common annotation tools. We thus identified an average of additional 1–5 variants per patient which summed up to the previous identified pathogenic variants (Fig 1, S3 Table).

Known common susceptibility/protective variants analysis

We then checked the cohort for known non rare variants classified as either "pathogenic" or "protective" in ClinVar database and related to viral infection. Variants in six different genes matched the term of "viral infection" and "pathogenic" according to ClinVar (Fig 1). Overall, a mean of 3 genes with "pathogenic" common variants involved in viral infection susceptibility were present (Fig 1).

Among the common protective variants, we list as example three variants which confer protection to Human Immunodeficiency Virus (HIV), the first two, and leprosy, the third one: a

CCR2 variant (rs1799864) identified in 8 patients, a *CCR5* (rs1800940) in one patient and a *TLR1* variant (rs5743618) in 26 patients (not shown). A *IL4R* variant (rs1805015) associated with HIV slow progression was present in 8 patients (not shown).

Candidate gene overview

Although not identified by unbiased collapsing gene analysis a number of obvious candidate genes were specifically analyzed. First, we noticed that SARS-CoV-2 receptor, ACE2 protein is preserved in the cohort, only a silent mutation V749V being present in 2 males and 2 heterozygous females. This is in line with our previous suggestion that either rare variants or polymorphisms may impact infectivity [10]. The *IFITM3* polymorphism (rs12252) was found in heterozygosity in 4 patients as expected by frequency. Eight patients had heterozygous missense mutations in *CFTR* gene reported as VUS/mild variants, 7 / 8 being among the more severely affected patients.

Discussion

In this study, we present a cohort of 35 COVID-19 patients admitted between April and May 2020 to the University Hospital of Siena who were clinically characterized by a team of 29 MDs belonging to 7 different specialties. As expected, the majority of hospitalized patients are males, confirming previously published data reporting a predominance of males among the most severe COVID-19 affected patients [11]. Lung imaging involvement, evaluated through a modified lung imaging grading system, did not completely correlate with respiratory impairment since among the 13 patients who required mechanical ventilation (group 1 and 2), grading was either moderate (10) or mild (3). In line with our previous data, lymphocyte subset immunophenotyping revealed a decrease in the total number of CD4 and NK cells count, especially in the most severe patients [12]. Laboratory tests revealed a multiple-organ involvement, confirming that COVID-19 is a systemic disease rather than just a lung disorder (Fig 1). We thus propose that only a detailed clinical characterization can allow to disentangle the complex relationship between genes and signs/symptoms.

In order to test the hypothesis that the COVID-19 susceptibility is due to one or more genes in common among patients, we used the gene burden test to compare the rate of disrupting mutations per gene. This test has already been successfully applied to discover susceptibility genes for Respiratory Syncytial Virus infection [13]. We identified 2 genes whose damage represents a susceptibility factor. Mutations in *PRKRA* (protein kinase activator A, alias PACT; OMIM# *603424), a protein kinase activated by viral double-stranded RNA may impair the down-stream IFN-mediated immune response [14, 15]. Mutations in *LAPTM4B* (Lysosomal Protein Transmembrane 4 Beta) gene, may impair endosomal network, eventually compromising productive viral infection [16, 17].

We then identified 2 genes whose damage represents a protective factor: *OR4C5* and *NDUFAF7*. *OR4C5* is a "resurrected" pseudogene, known to be non functioning in half of the European population, with a frequency of inactive allele of 0.62 in Asians, 0.48 in Europeans and 0.16 in Africans [18, 19]. Expression of the "resurrected" pseudogene *OR4C5* may help in triggering the natural immunity leading to virus and cell death [20, 21]. It is interesting to note that protein atlas shows *OR4C5* protein expression in the liver without the corresponding mRNA expression (www.proteinatlas.org) suggesting that *OR4C5* reaches the liver through nerve terminals [22]. If this is the case, those individuals expressing the resurrected *OR4C5* gene may have more triggers of innate immunity and subsequently higher liver damage, in agreement with the putative expression of *OR4C5* (white boxes) in patients with liver impairment (Fig 1).

Previous studies reported a prevalence of olfactory disorders in COVID-19 population ranging from 5% to 98%. A recent meta-analysis of 10 studies demonstrated a 52.73% prevalence for smell dysfunction in COVID-19 subjects [23]. In our population, only 3/35 (8.6%) subjects reported olfactory disorders. Both the limited sample size and the characteristic of the population (severely affected hospitalized subjects) could explain this result. However, a report focusing on smell dysfunction in severely affected hospitalized subjects reported a prevalence of 23.7% among 59 patients [24].

We explored the hypothesis that each patient could have one unique combination of rare pathogenic/highly relevant variants related for different reasons to infection susceptibility [9] (Fig 1): G6PD-deficient cells are more susceptible to several viruses including coronavirus and have down-regulated innate immunity (in line with the observed very low levels of IL-6) (Fig 1) [25]; *ZEB1*-linked corneal dystrophy, known to function in immune cells, and playing an important role in establishing both the effector response and future immunity in response to pathogens [26]; *TGFBI* mutations (associated with corneal dystrophy); *ABCC6* gene mutations (associated with pseudoxanthoma elasticum); likely hypomorphic mutations in *CHD7* or *COL5A1/2* variants, playing a role as modulators of immune cells activity and/or response to infections [27–34]; *ADAR*, involved in viral RNA editing; *CLEC4M*, an alternative receptor for SARS-CoV [35] *HCRTR1/2*, receptors of Hypocretin, important in the regulation of fatigue during infections [36]; *FURIN*, a serine protease that cleaves the SARS-Cov-2 minor capsid protein important for ACE2 contact and viral entry into the host cells [37, 38].

Finally, interesting rare variants have been identified in NitricOxide synthase *NOS3* and Opioid receptor *OPRM1*. Opioid ligands may regulate the expression of chemokines and chemokine receptors [39]. NitricOxide (NO), mainly produced by epithelial and white blood cells (iNOS) and to a lesser extent by endothelial cells (eNOS), is able to significantly reduce viral infection and replication of SARS-CoV in normal condition through two distinct mechanisms: impairment of the fusion between the spike protein and its receptor ACE2, and reduction of viral RNA production [40]. Mutations in NO synthase may disrupt one or both the above reported functions and clinical trials are ongoing to evaluate the effectiveness of inhaled NO in COVID-19 patients [41, 42].

Several rare variants in Interleukins (*ILs*) and Interleukins receptors (*ILRs*) are found. Interleukins are crucial in modulating immune response against all types of infective agents. The variants reported in this study include different interleukins that are not specifically involved in the defense against virus but are critical in balancing both innate and specific adaptive immune response (Fig 1).

Furthemore, we identified common "pathogenic" variants in genes known to be linked to viral infection, such as *MBL2*, *IRGM* and *SAA1*, and/or specific organ damage as *PRSS1*. Polymorphisms in *PRSS1*, a serine protease secreted from the pancreas, are associated with autosomal dominant hereditary pancreatitis (OMIM#167800) [43]. Polymorphisms in *MBL2*, a mannose-binding lectin secreted by the liver, cause increased susceptibility to infections, possibly due to a negative impact on the ability to mount an immune response [44, 45]. Polymorphisms in *IRGM* may lead to impairment of autophagy which in turn controls innate and adaptive immunity [46, 47]. *SAA1*, encoding the serum amyloid A (SAA) protein, is an apolipoprotein reactant, mainly produced by hepatocytes and regulated from inflammatory cytokines. In patients with chronic inflammatory diseases, the SAA cleavage product, Amyloid protein A (AA), is deposited systemically in vital organs including liver, spleen and kidneys, causing amyloidosis [48].

For the last above reported genes and pathogenic variants or predicted variants relevant for infection, a statistically significant difference in variant's frequency was not found between cases and controls looking at either the single variant or the single gene, as a burden effect of

variants. However, as depicted in the overall Fig 1, we could hypothesize a combined model in which common susceptibility genes will sum to less common or private susceptibility variants. A specific combination of these 2 categories may determine type (organotropism) and severity of the disease.

Our observations related to the huge amount of data, both on phenome and genome sides, and represented in Fig 1, could also lay the bases for association rule mining approaches. Artificial intelligence techniques based on pattern recognition may discover an intelligible picture which appears blurred at present.

We know that a possible limitation of this study is the heterogeneity of patients and controls, which are not matched for gender, major comorbidities and other clinical characteristics. For this reason, further analyses in a larger cohort of samples are mandatory in order to test this hypothesis of a combined model for COVID-19 susceptibility with a number of common susceptibility genes which represent the fertile background in which additional private, rare or low frequency mutations confer to the host the most favorable environment for virus growth and organ damage.

Methods

Patients clinical data and samples collection

The GEN-COVID study was approved by the University Hospital of Siena Ethical Review Board (Prot n. 16929, dated March 16, 2020). Thirty-five patients admitted to the University Hospital in Siena, Italy, from April 7 to May 7, 2020 were recruited. WES data of these 35 patients were compared with those of 150 controls (the Siena cohort of the Network of Italian Genomes NIG http://www.nig.cineca.it). Patients have a mean age of 64 years with a Standard Deviation (SD) of 14.3 while the controls have a mean age of 46 years with a SD of 9.5. The percentage of males (M) and females (F) in patients is 68.5% and 31.4% respectively, while in controls is 51% and 49% respectively. The patients are clustered into four qualitative severity groups depending on the respiratory impairment and the need for ventilation: high care intensity group (those requiring invasive ventilation), intermediate care intensity group (those requiring non invasive ventilation i.e. CPAP and BiPAP, and high-flows oxygen therapy), low care intensity group (those requiring conventional oxygen therapy) and very low care intensity group (those not requiring oxygen therapy) (groups 1–4 in Table 1 and different colors in Fig 1).

Peripheral blood samples in EDTA-containing tubes and detailed clinical data were collected. All these data were inserted in a section dedicated to COVID-19 of the established and certified Biobank and Registry of the Medical Genetics Unit of the Hospital. An example of the Clinical questionnaire is illustrated in S1 Fig.

Each patient was assigned a continuous quantitative respiratory score, the PaO2/FiO2 ratio (normal values >300) (P/F), as the worst value during the hospitalization.

Patients were also assigned a lung imaging grading according to X-Rays and CT scans. In particular, lung involvement was scored through imaging at the time of admission and during hospitalization (worst score), annotating the chest X-Ray (CXR) score (in 34 patients) and CT score in 1 patient for whom X-Rays were not available. To obtain the score (from 0 to 28) each CXR was divided in four quadrant (right upper, right lower, left upper and left lower) and for each quadrant the presence of consolidation (0 = no consolidation; 1 <50%, 2>50%), ground glass opacities (GGOs: 0 = no GGOs, 1<50%, 2 >50%), reticulation (0 = no GGOs, 1<50%, 2 >50%) and pleural effusion on left or right side (0 = no, 1 = minimal; 2 = large) were recorded. The same score was applied for CT (1 patient).

For each patient, the presence of hyposmia and hypogeusia was also investigated through otolaryngology examination, Burghart sniffin' sticks [49] and a visual analog scale (VAS). Whenever the sign was present, a score ranging from 0 to 10 was assigned to each patient using VAS where 0 means the best sense of smell and 10 represents the absence of smell sensation [50].

The presence of hepatic involvement was defined on the basis of a clear hepatic enzymes elevation as glutamic pyruvic transaminase (ALT) and glutamic oxaloacetic transaminase (AST) both higher than 40 UI/l. Pancreatic involvement was considered on the basis of an increase of pancreatic enzymes as pancreatic amylase higher than 53 UI/l and lipase higher than 60UI/l. Heart involvement was defined on the basis of one or more of the following abnormal data: Troponin T (>15 ng/L), indicative of ischemic disorder; NT-proBNP (M >88; F >153 pg/ml), indicative of heart failure and arrhythmias (indicative of electric disorder). Kidney involvement was defined in the presence of a creatinine value higher than 1,20 mg/dl in males and higher than 1,10 mg/dl in females (Fig 1).

Whole exome sequencing analysis

Genomic DNA was extracted from peripheral blood using the MagCore® Genomic DNA Whole Blood kit (RBC Biosciences) according to manufacturer's protocol. Whole exome sequencing analysis was performed on Illumina NovaSeq 6000 system (Illumina, San Diego, CA, USA). DNA fragments were hybridized and captured by Illumina Exome Panel (Illumina) according to manufacturer's protocol. The libraries were tested for enrichment by qPCR, and the size distribution and concentration were determined using an Agilent Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA, USA). The Novaseq 6000 platform (Illumina), along with 150 bp paired-end reads, was used for sequencing of DNA.

Genetic data analysis

Reads were mapped to the hg19 reference genome by the Burrow-Wheeler aligner BWA [51]. Variants calling was performed according to the GATK4 best practice guidelines [52]. Namely, duplicates were first removed by *MarkDuplicates*, and base qualities were recalibrated using *BaseRecalibration* and *ApplyBQSR*. *HaplotypeCaller* was used to calculate Genomic VCF files for each sample, which were then used for multi-sample calling by *GenomicDBImport* and *GenotypeGVCF*. In order to improve the specificity-sensitivity balance, variants quality scores were calculated by *VariantRecalibrator* and *ApplyVQSR*. Variants were annotated by ANNOVAR [53], and with the number of articles answering the query "gene_name AND viral infection" in Pubmed, where gene_name is the name of the gene affected by the variant.

In order to identify candidate genes according to the Mendelian-like model, rare variants were filtered by a prioritization approach. We used the ExAC database (http://exac.broadinstitute.org/), in particular the ExAC_NFE reported frequency to filter variants according to a minor allele frequency < 0.01. Synonymous, intronic and non-coding variants were excluded from the analysis. Mutation disease database ClinVar (ncbi.nlm.nih.gov/clinvar/) was used to identify previous pathogenicity classifications and variants reported as likely benign/benign were discarded. Filtering and prioritization of variants was completed using the CADD_Phred pathogenicity prediction tool. Finally, we selected genes involved in infection susceptibility using the term "viral infection" as Pubmed database search.

In order to identify genes with a different prevalence of functionally relevant variants between COVID-19 patients and control samples, the following score was calculated:

$$x_j = \sum_{i=1}^{K} w_i x_{i,j}, \quad (1)$$

Where w_i is a weight associated with the *i-th* variant; and $x\_(i, j)$ is equal to 0 if the variant is not present in sample *j*, 1 if sample *j* has the variant in heterozygous state, and 2 if sample *j* has the variant is homozygous state. The weight w_i was assumed equal to the DANN score of the variant [54], which provides an estimate of the likelihood that the variant has deleterious functional effects (i.e. variants more likely to have a functional effect contribute more to the score). The sum in equation (1) was performed over all the variants in the gene where the DANN score was available. Genes with less than 5 annotated variants were discarded from the analysis. The scores calculated by equation (1) were ranked for all the samples, and the sum of the ranking for the COVID-19 samples, named r_{COVID}, was calculated. Then, sample labels were permuted 10.000 times, and these permutations were used to estimate the average value and the standard deviation of r_{COVID} under the null-hypothesis. The p-value was calculated assuming a normal distribution for the sum of the ranking [55]. Moreover, we performed an additional more stringent quality check of genetic variants in the selected genes in order to remove calling artifacts that skipped the previous quality control.

Supporting information

S1 Fig. Clinical Data Questionnaire. The Questionnaire includes five different categories of data: Patient personal anamnesis and family history, Diagnostic Information, Laboratory Tests, Therapy and Complications. Clinical data were collected in detail for all COVID-19 patients.
(TIF)

S1 Table. List of genes conferring COVID-19 susceptibility identified with the gene burden test analysis. Genes harboring deleterious mutations with statistically significant higher frequency in control than in COVID-19 patients are ordered based on p-value deriving from gene burden test analysis. The p-value adjusted is provided after Bonferroni correction.
(XLSX)

S2 Table. List of COVID-19 protective genes identified with the gene burden test analysis. Genes harboring deleterious mutations with statistically significant higher frequency in COVID-19 patients than in control are ordered based on p-value deriving from gene burden test analysis. The p-value adjusted is provided after Bonferroni correction.
(XLSX)

S3 Table. Rare variants identified in patients cohort. Rare variants identified in COVID-19 patients according to the Mendelian-like model are reported (see Methods *section*).
(XLSX)

Acknowledgments

This study is part of GEN-COVID, https://sites.google.com/dbm.unisi.it/gen-covid the Italian multicenter study aimed to identify the COVID-19 host genetic bases The *Genetic and COVID-19 Biobank of Siena*, member of BBMRI-IT, of Telethon Network of Genetic Biobanks (project no. GTB18001), of EuroBioBank, and of D-Connect, provided us with specimens. We thank the CINECA consortium for providing computational resources and Network for

Italian Genomes NIG http://www.nig.cineca.it. We thank private donors' support to A.R. (Department of Medical Biotechnologies, University of Siena) for the COVID-19 host genetics research project (D.L n.18 of March 17th 2020).

GEN-COVID Multicenter Study (composition at May 22, 2020, the representative of the GEN-COVID multicenter study is Prof. Francesca Mari email: francesca.mari@unisi.it**)**

Gabriella Doddato[1], Susanna Croci[1], Laura Di Sarno[1], Andrea Tommasi[1,2], Sergio Daga[1], Maria Palmieri[1], Massimiliano Fabbiani[5], Barbara Rossetti[5], Giacomo Zanelli[3,5], Paolo Cameli[6], David Bennett[6], Simona Marcantonio[7], Sabino Scolletta[7], Federico Franchi[7], Luca Cantarini[9], Bruno Frediani[9], Danilo Tacconi[10], Chiara Spertilli[10], Marco Feri[11], Alice Donati[11], Raffaele Scala[12], Luca Guidelli[12], Agostino Ognibene[13], Genni Spargi[14], Marta Corridi[14], Cesira Nencioni[15], Leonardo Croci[15], Gian Piero Caldarelli[16], Maurizio Spagnesi[17], Paolo Piacentini[17], Anna Canaccini[18], Agnese Verzuri[18], Valentina Anemoli[18], Massimo Vaghi[21], Antonella D'Arminio Monforte[22], Esther Merlini[22], Mario Umberto Mondelli[23,24], Stefania Mantovani[23], Serena Ludovisi[24], Massimo Girardis[25], Sophie Venturelli[25], Andrea Cossarizza[26], Andrea Antinori[27], Alessandra Vergori[27], Stefano Rusconi[28,29], Matteo Siano[28,29], Arianna Gabrieli[29], Daniela Francisci[30,31], Elisabetta Schiaroli[30], Pier Giorgio Scotton[32], Francesca Andretta[32], Sandro Panese[33], Renzo Scaggiante[34], Saverio Giuseppe Parisi[35], Francesco Castelli[36], Maria Eugenia Quiros Roldan[36], Paola Magro[36], Cristina Minardi[36], Matteo Della Monica[37], Carmelo Piscopo[37], Mario Capasso[38,39,40], Massimo Carella[41], Marco Castori[41], Giuseppe Merla[41], Filippo Aucella[42], Pamela Raggi[43], Matteo Bassetti[44,45], Antonio Di Biagio[45], Maurizio Sanguinetti[46,47], Luca Masucci[46,47], Chiara Gabbi[19], Serafina Valente[18], Susanna Guerrini[8], Elisa Frullanti[1], Ilaria Meloni[1], Maria Antonietta Mencarelli[2], Caterina Lo Rizzo[2], Anna Maria Pinto[2]

10) Department of Specialized and Internal Medicine, Infectious Diseases Unit, San Donato Hospital Arezzo, Italy

11) Department of Emergency, Anesthesia Unit, San Donato Hospital, Arezzo, Italy

12) Department of Specialized and Internal Medicine, Pneumology Unit and UTIP, San Donato Hospital, Arezzo, Italy

13) Clinical Chemical Analysis Laboratory, San Donato Hospital, Arezzo, Italy

14) Department of Emergency, Anesthesia Unit, Misericordia Hospital, Grosseto, Italy

15) Department of Specialized and Internal Medicine, Infectious Diseases Unit, Misericordia Hospital, Grosseto, Italy

16) Clinical Chemical Analysis Laboratory, Misericordia Hospital, Grosseto, Italy

17) Department of Prevention, Azienda USL Toscana Sud Est, Italy

18) Territorial Scientific Technician Department, Azienda USL Toscana Sud Est, Italy

19) Independent Scientist, Milan, Italy

20) Department of Cardiovascular Diseases, University of Siena, Italy

21) Chirurgia Vascolare, Ospedale Maggiore di Crema, Italy

22) Department of Health Sciences, Clinic of Infectious Diseases, ASST Santi Paolo e Carlo, University of Milan, Italy

23) Division of Infectious Diseases and Immunology, Department of Medical Sciences and Infectious Diseases, Pavia, Italy.

24) Department of Internal Medicine and Therapeutics, University of Pavia, Italy

25) Department of Anesthesia and Intensive Care, University of Modena and Reggio Emilia, Modena, Italy

26) Department of Medical and Surgical Sciences for Children and Adults, University of Modena and Reggio Emilia, Modena, Italy

27) HIV/AIDS Department, National Institute for Infectious Diseases, IRCCS, Lazzaro Spallanzani, Rome, Italy

28) III Infectious Diseases Unit, ASST-FBF-Sacco, Milan, Italy
29) Department of Biomedical and Clinical Sciences Luigi Sacco, University of Milan, Milan, Italy
30) Infectious Diseases Clinic, Department of Medicine 2, Azienda Ospedaliera di Perugia and University of Perugia, Santa Maria Hospital, Perugia, Italy
31) Infectious Diseases Clinic, "Santa Maria" Hospital, University of Perugia, Perugia, Italy
32) Department of Infectious Diseases, Treviso Hospital, Local Health Unit 2 Marca Trevigiana, Treviso, Italy
33) Infectious Diseases Department, Ospedale Civile "SS. Giovanni e Paolo", Venice, Italy
34) Infectious Diseases Clinic, ULSS1, Belluno, Italy
35) Department of Molecular Medicine, University of Padova, Italy
36) Department of Infectious and Tropical Diseases, University of Brescia and ASST Spedali Civili Hospital, Brescia, Italy.
37) Medical Genetics and Laboratory of Medical Genetics Unit, A.O.R.N. "Antonio Cardarelli", Naples, Italy.
38) Department of Molecular Medicine and Medical Biotechnology, University of Naples Federico II, Naples, Italy.
39) CEINGE Biotecnologie Avanzate, Naples, Italy
40) IRCCS SDN, Naples, Italy.
41) Division of Medical Genetics, Fondazione IRCCS Casa Sollievo della Sofferenza Hospital, San Giovanni Rotondo, Italy.
42) Department of Nephrology and Dialysis, Fondazione IRCCS Casa Sollievo della Sofferenza Hospital, San Giovanni Rotondo, Italy.
43) Department of Medical Sciences, Fondazione IRCCS Casa Sollievo della Sofferenza Hospital, San Giovanni Rotondo, Italy.
44) Department of Health Sciences, University of Genova, Genova, Italy.
45) Infectious Diseases Clinic, Policlinico San Martino Hospital, IRCCS for Cancer Research Genova, Italy.
46) Microbiology, Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Catholic University of Medicine, Rome, Italy.
47) Department of Laboratory Sciences and Infectious Diseases, Fondazione Policlinico Universitario A. Gemelli IRCCS, Rome, Italy.

Author Contributions

Conceptualization: Annarita Giliberti, Alessandra Renieri, Francesca Mari.

Data curation: Elisa Benetti, Chiara Fallerini, Rossella Tita, Simone Furini.

Formal analysis: Elisa Benetti, Simone Furini.

Investigation: Annarita Giliberti, Arianna Emiliozzi, Floriana Valentino, Laura Bergantini, Federico Anedda, Sara Amitrano, Edoardo Conticini, Miriana d'Alessandro, Francesca Fava, Simona Marcantonio, Margherita Baldassarri, Mirella Bruttini, Maria Antonietta Mazzei, Francesca Montagnani, Marco Mandalà, Elena Bargagli, Alessandra Renieri, Francesca Mari.

Project administration: Alessandra Renieri, Francesca Mari.

Software: Simone Furini.

Writing – original draft: Elisa Benetti, Alessandra Renieri, Francesca Mari.

References

1. Dennison Himmelfarb CR, Baptiste D. Coronavirus Disease (COVID-19). J Cardiovasc Nurs. 2020; Publish Ah. https://doi.org/10.1097/jcn.0000000000000710 PMID: 32384299
2. Guan W, Ni Z, Hu Y, Liang W, Ou C, He J, et al. Clinical characteristics of coronavirus disease 2019 in China. N Engl J Med. 2020. https://doi.org/10.1056/NEJMoa2002032 PMID: 32109013
3. Wu Z, McGoogan JM. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China. JAMA. 2020. https://doi.org/10.1001/jama.2020.2648 PMID: 32091533
4. Liu R, Paxton WA, Choe S, Ceradini D, Martin SR, Horuk R, et al. Homozygous defect in HIV-1 coreceptor accounts for resistance of some multiply-exposed individuals to HIV-1 infection. Cell. 1996. https://doi.org/10.1016/s0092-8674(00)80110-5 PMID: 8756719
5. Woziwodzka A, Rybicka M, Sznarkowska A, Romanowski T, Dręczewski M, Stalke P, et al. TNF-α polymorphisms affect persistence and progression of HBV infection. Mol Genet Genomic Med. 2019. https://doi.org/10.1002/mgg3.935 PMID: 31441603
6. Tian T, Huang P, Wu J, Wang C, Fan H, Zhang Y, et al. CD40 polymorphisms were associated with HCV infection susceptibility among Chinese population. BMC Infect Dis. 2019. https://doi.org/10.1186/s12879-019-4482-5 PMID: 31615434
7. Nogales A, Dediego ML. Host single nucleotide polymorphisms modulating influenza a virus disease in humans. Pathogens. 2019. https://doi.org/10.3390/pathogens8040168 PMID: 31574965
8. Williams FMK, Freydin M, Mangino M, Couvreur S, Visconti A, Bowyer RCE, et al. Self-reported symptoms of covid-19 including symptoms most predictive of SARS-CoV-2 infection, are heritable. MedRxiv. 2020.
9. Satterstrom FK, Kosmicki JA, Wang J, Breen MS, De Rubeis S, An JY, et al. Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. Cell. 2020. https://doi.org/10.1016/j.cell.2019.12.036 PMID: 31981491
10. Benetti E, Tita R, Spiga O, Ciolfi A, Birolo G, Bruselles A, et al. ACE2 gene variants may underlie interindividual variability and susceptibility to COVID-19 in the Italian population. Eur J Hum Genet. 2020. https://doi.org/10.1038/s41431-020-0691-z PMID: 32681121
11. Cai H. Sex difference and smoking predisposition in patients with COVID-19. The Lancet Respiratory Medicine. 2020. https://doi.org/10.1016/S2213-2600(20)30117-X PMID: 32171067
12. D'alessandro M, Bennett D, Montagnani F, Cameli P, Perrone A, Bergantini L, et al. Peripheral lymphocyte subset monitoring in COVID19 patients: a prospective Italian real-life case series. Minerva Med. 2020. https://doi.org/10.23736/S0026-4806.20.06638-0 PMID: 32407057
13. Salas A, Pardo-Seco J, Cebey-López M, Gómez-Carballa A, Obando-Pacheco P, Rivero-Calle I, et al. Whole Exome Sequencing reveals new candidate genes in host genomic susceptibility to Respiratory Syncytial Virus Disease. Sci Rep. 2017. https://doi.org/10.1038/s41598-017-15752-4 PMID: 29162850
14. Chan CP, Yuen CK, Cheung PHH, Fung SY, Lui PY, Chen H, et al. Antiviral activity of double-stranded RNA-binding protein PACT against influenza A virus mediated via suppression of viral RNA polymerase. FASEB J. 2018; 32: 4380–4393. https://doi.org/10.1096/fj.201701361R PMID: 29513570
15. Miyamoto M, Komuro A. PACT is required for MDA5-mediated immunoresponses triggered by Cardiovirus infection via interaction with LGP2. Biochem Biophys Res Commun. 2017. https://doi.org/10.1016/j.bbrc.2017.10.048 PMID: 29032202
16. Iwamoto M, Saso W, Sugiyama R, Ishii K, Ohki M, Nagamori S, et al. Epidermal growth factor receptor is a host-entry cofactor triggering hepatitis B virus internalization. Proc Natl Acad Sci U S A. 2019. https://doi.org/10.1073/pnas.1811064116 PMID: 30952782
17. Tan X, Sun Y, Thapa N, Liao Y, Hedman AC, Anderson RA. LAPTM4B is a PtdIns(4,5)P 2 effector that regulates EGFR signaling, lysosomal sorting, and degradation. EMBO J. 2015. https://doi.org/10.15252/embj.201489425 PMID: 25588945
18. Olender T, Waszak SM, Viavant M, Khen M, Ben-Asher E, Reyes A, et al. Personal receptor repertoires: olfaction as a model. BMC Genomics. 2012. https://doi.org/10.1186/1471-2164-13-414 PMID: 22908908
19. Waszak SM, Hasin Y, Zichner T, Olender T, Keydar I, Khen M, et al. Systematic inference of copy-number genotypes from personal genome sequencing data reveals extensive olfactory receptor gene content diversity. PLoS Comput Biol. 2010. https://doi.org/10.1371/journal.pcbi.1000988 PMID: 21085617
20. Durrant DM, Ghosh S, Klein RS. The Olfactory Bulb: An Immunosensory Effector Organ during Neurotropic Viral Infections. ACS Chemical Neuroscience. 2016. https://doi.org/10.1021/acschemneuro.6b00043 PMID: 27058872
21. Mori I, Goshima F, Imai Y, Kohsaka S, Sugiyama T, Yoshida T, et al. Olfactory receptor neurons prevent disseminations of neurovirulent influenza A virus into the brain by undergoing virus-induced

apoptosis. J Gen Virol. 2002; 83: 2109–2116. https://doi.org/10.1099/0022-1317-83-9-2109 PMID: 12185263

22. Streba LAM, Vere CC, Ionescu AG, Streba CT, Rogoveanu I. Role of intrahepatic innervation in regulating the activity of liver cells. World Journal of Hepatology. 2014. https://doi.org/10.4254/wjh.v6.i3.137 PMID: 24672643

23. Tong JY, Wong A, Zhu D, Fastenberg JH, Tham T. The Prevalence of Olfactory and Gustatory Dysfunction in COVID-19 Patients: A Systematic Review and Meta-analysis. Otolaryngology—Head and Neck Surgery (United States). 2020. https://doi.org/10.1177/0194599820926473 PMID: 32369429

24. Giacomelli A, Pezzati L, Conti F, Bernacchia D, Siano M, Oreni L, et al. Self-reported olfactory and taste disorders in SARS-CoV-2 patients: a cross-sectional study. Clin Infect Dis. 2020. https://doi.org/10.1093/cid/ciaa330 PMID: 32215618

25. Wu YH, Tseng CP, Cheng ML, Ho HY, Shih SR, Chiu DTY. Glucose-6-phosphate dehydrogenase deficiency enhances human coronavirus 229E infection. J Infect Dis. 2008. https://doi.org/10.1086/528377 PMID: 18269318

26. Guan T, Dominguez CX, Amezquita RA, Laidlaw BJ, Cheng J, Henao-Mejia J, et al. ZEB1, ZEB2, and the miR-200 family form a counterregulatory network to regulate CD8+ T cell fates. J Exp Med. 2018. https://doi.org/10.1084/jem.20171352 PMID: 29449309

27. Klamer SE, Dorland YL, Kleijer M, Geerts D, Lento WE, Van Der Schoot CE, et al. TGFBI expressed by bone marrow niche cells and hematopoietic stem and progenitor cells regulates hematopoiesis. Stem Cells Dev. 2018. https://doi.org/10.1089/scd.2018.0124 PMID: 30084753

28. Ebersole JL, Peyyala R, Gonzalez OA. Biofilm-induced profiles of immune response gene expression by oral epithelial cells. Mol Oral Microbiol. 2019. https://doi.org/10.1111/omi.12251 PMID: 30407731

29. Marton J, Albert D, Wiltshire SA, Park R, Bergen A, Qureshi S, et al. Cyclosporine a treatment inhibits Abcc6-dependent cardiac necrosis and calcification following coxsackievirus B3 infection in mice. PLoS One. 2015. https://doi.org/10.1371/journal.pone.0138222 PMID: 26375467

30. Janssen N, Bergman JEH, Swertz MA, Tranebjaerg L, Lodahl M, Schoots J, et al. Mutation update on the CHD7 gene involved in CHARGE syndrome. Human Mutation. 2012. https://doi.org/10.1002/humu.22086 PMID: 22461308

31. Theodoropoulos DS, Theodoropoulos GA, Edwards BM, Kileny PR, Van Riper LA. Immune deficiency and hearing loss in CHARGE association [3]. Pediatrics. 2003. https://doi.org/10.1542/peds.111.3.711-a PMID: 12612267

32. Gennery AR, Slatter MA, Rice J, Hoefsloot LH, Barge D, McLean-Tooke A, et al. Mutations in CHD7 in patients with CHARGE syndrome cause T-B + natural killer cell + severe combined immune deficiency and may cause Omenn-like syndrome. Clin Exp Immunol. 2008. https://doi.org/10.1111/j.1365-2249.2008.03681.x PMID: 18505430

33. Randall V, McCue K, Roberts C, Kyriakopoulou V, Beddow S, Barrett AN, et al. Great vessel development requires biallelic expression of Chd7 and Tbx1 in pharyngeal ectoderm in mice. J Clin Invest. 2009. https://doi.org/10.1172/JCI37561 PMID: 19855134

34. Zhetkenev S, Khassan A, Khamzina A, Issanov A, Crape B, Akilzhanova A, et al. Association of rs12722 COL5A1 with Pulmonary Tuberculosis infection: a preliminary case-control study in a Kazakhstani population. 2019; 2017: 19008995. https://doi.org/10.1101/19008995

35. Chan VSF, Chan KYK, Chen Y, Poon LLM, Cheung ANY, Zheng B, et al. Homozygous L-SIGN (CLEC4M) plays a protective role in SARS coronavirus infection. Nat Genet. 2006. https://doi.org/10.1038/ng1698 PMID: 16369534

36. Zhan S, Cai GQ, Zheng A, Wang Y, Jia J, Fang H, et al. Tumor necrosis factor-alpha regulates the Hypocretin system via mRNA degradation and ubiquitination. Biochim Biophys Acta—Mol Basis Dis. 2011. https://doi.org/10.1016/j.bbadis.2010.11.003 PMID: 21094253

37. Braun E, Hotter D, Koepke L, Zech F, Groß R, Sparrer KMJ, et al. Guanylate-Binding Proteins 2 and 5 Exert Broad Antiviral Activity by Inhibiting Furin-Mediated Processing of Viral Envelope Proteins. Cell Rep. 2019. https://doi.org/10.1016/j.celrep.2019.04.063 PMID: 31091448

38. Shang J, Wan Y, Luo C, Ye G, Geng Q, Auerbach A, et al. Cell entry mechanisms of SARS-CoV-2. Proc Natl Acad Sci U S A. 2020. https://doi.org/10.1073/pnas.2003138117 PMID: 32376634

39. Finley MJ, Happel CM, Kaminsky DE, Rogers TJ. Opioid and nociceptin receptors regulate cytokine and cytokine receptor expression. Cellular Immunology. 2008. https://doi.org/10.1016/j.cellimm.2007.09.008 PMID: 18279847

40. Åkerström S, Gunalan V, Keng CT, Tan YJ, Mirazimi A. Dual effect of nitric oxide on SARS-CoV replication: Viral RNA production and palmitoylation of the S protein are affected. Virology. 2009. https://doi.org/10.1016/j.virol.2009.09.007 PMID: 19800091

41. Åkerström S, Mousavi-Jazi M, Klingström J, Leijon M, Lundkvist Å, Mirazimi A. Nitric Oxide Inhibits the Replication Cycle of Severe Acute Respiratory Syndrome Coronavirus. J Virol. 2005. https://doi.org/10.1128/JVI.79.3.1966-1969.2005 PMID: 15650225

42. Zamanian RT, Pollack C V., Gentile MA, Rashid M, Fox JC, Mahaffey KW, et al. Outpatient inhaled nitric oxide in a patient with vasoreactive idiopathic pulmonary arterial hypertension and COVID-19 infection. American Journal of Respiratory and Critical Care Medicine. 2020. https://doi.org/10.1164/rccm.202004-0937LE PMID: 32369396

43. Teich N, Nemoda Z, Köhler H, Heinritz W, Mössner J, Keim V, et al. Gene conversion between functional trypsinogen genes PRSS1 and PRSS2 associated with chronic pancreatitis in a six-year-old girl. Hum Mutat. 2005. https://doi.org/10.1002/humu.20148 PMID: 15776435

44. Thio CL, Mosbruger T, Astemborski J, Greer S, Kirk GD, O'Brien SJ, et al. Mannose Binding Lectin Genotypes Influence Recovery from Hepatitis B Virus Infection. J Virol. 2005. https://doi.org/10.1128/JVI.79.14.9192-9196.2005 PMID: 15994813

45. Dean MM, Flower RL, Eisen DP, Minchinton RM, Hart DNJ, Vuckovic S. Mannose-binding lectin deficiency influences innate and antigen-presenting functions of blood myeloid dendritic cells. Immunology. 2011. https://doi.org/10.1111/j.1365-2567.2010.03365.x PMID: 21091907

46. Singh SB, Davis AS, Taylor GA, Deretic V. Human IRGM induces autophagy to eliminate intracellular mycobacteria. Science (80-). 2006. https://doi.org/10.1126/science.1129577 PMID: 16888103

47. Rufini S, Ciccacci C, Di Fusco D, Ruffa A, Pallone F, Novelli G, et al. Autophagy and inflammatory bowel disease: Association between variants of the autophagy-related IRGM gene and susceptibility to Crohn's disease. Dig Liver Dis. 2015. https://doi.org/10.1016/j.dld.2015.05.012 PMID: 26066377

48. Zhang Y, Zhang J, Sheng H, Li H, Wang R. Acute phase reactant serum amyloid A in inflammation and other diseases. Advances in Clinical Chemistry. 2019. https://doi.org/10.1016/bs.acc.2019.01.002 PMID: 31122611

49. Oleszkiewicz A, Schriever VA, Croy I, Hähner A, Hummel T. Updated Sniffin' Sticks normative data based on an extended sample of 9139 subjects. Eur Arch Oto-Rhino-Laryngology. 2019. https://doi.org/10.1007/s00405-018-5248-1 PMID: 30554358

50. Klimek L, Bergmann KC, Biedermann T, Bousquet J, Hellings P, Jung K, et al. Visual analogue scales (VAS)—Measuring instruments for the documentation of symptoms and therapy monitoring in case of allergic rhinitis in everyday health care. Allergo J. 2017. https://doi.org/10.1007/s40629-016-0006-7 PMID: 28217433

51. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010. https://doi.org/10.1093/bioinformatics/btp698 PMID: 20080505

52. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. 2017. https://doi.org/10.1101/201178

53. Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010. https://doi.org/10.1093/nar/gkq603 PMID: 20601685

54. Quang D, Chen Y, Xie X. DANN: A deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. 2015. https://doi.org/10.1093/bioinformatics/btu703 PMID: 25338716

55. Dering C, Hemmelmann C, Pugh E, Ziegler A. Statistical analysis of rare sequence variants: An overview of collapsing methods. Genet Epidemiol. 2011. https://doi.org/10.1002/gepi.20643 PMID: 22128052

5. Shorter androgen receptor polyQ alleles protect against life-threatening COVID-19 disease in European males

Male sex has been reported as a risk factor for worse COVID-19 outcome even if men and women are similarly infected by the virus. In this study we aim to evaluate if the variability in COVID-19 severity among males and females may be explained by differences in the host genome.

In this chapter, we report the first analysis carried out by exploiting the LASSO logistic regression on the genetic dataset of COVID-19. This is our first application of a synthetic representation of genetic variants in a machine learning model. To deepen the sex differences in COVID-19, we evaluate the potential impact of poly-amino acids repeat polymorphisms, via the Boolean feature of poly-amino acids triplet repeats (C_PR described in chapter 2, section 2.5.1). The polyQ tract of the Androgen Receptor (*AR*) gene resulted a key determinant [59].

EBioMedicine 65 (2021) 103246

Contents lists available at ScienceDirect

EBioMedicine

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

Research paper

Shorter androgen receptor polyQ alleles protect against life-threatening COVID-19 disease in European males

Margherita Baldassarri[a,b,1], Nicola Picchiotti[c,d,1], Francesca Fava[a,b,e], Chiara Fallerini[a,b], Elisa Benetti[b], Sergio Daga[a,b], Floriana Valentino[a,b], Gabriella Doddato[a,b], Simone Furini[b], Annarita Giliberti[a,b], Rossella Tita[e], Sara Amitrano[e], Mirella Bruttini[a,b,e], Susanna Croci[a,b], Ilaria Meloni[a,b], Anna Maria Pinto[e], Nicola Iuso[a,b], Chiara Gabbi[f], Francesca Sciarra[g], Mary Anna Venneri[g], Marco Gori[c,h], Maurizio Sanarico[i], Francis P. Crawley[j], Uberto Pagotto[k], Flaminia Fanelli[k], Marco Mezzullo[k], Elena Dominguez-Garrido[l], Laura Planas-Serra[m,n,o], Agatha Schlüter[m,n,o], Roger Colobran[p], Pere Soler-Palacin[q], Pablo Lapunzina[n,r], Jair Tenorio[n,r], Aurora Pujol[m,n,s], Maria Grazia Castagna[t], Marco Marcelli[u], Andrea M. Isidori[g], Alessandra Renieri[a,b,e,*], Elisa Frullanti[a,b,2], Francesca Mari[a,b,e,2], Spanish Covid HGE, GEN-COVID Multicenter Study

[a] Medical Genetics, University of Siena, Italy
[b] Med Biotech Hub and Competence Center, Department of Medical Biotechnologies, University of Siena, Italy
[c] University of Siena, DIISM-SAILAB, Siena, Italy
[d] Department of Mathematics, University of Pavia, Pavia, Italy
[e] Genetica Medica, Azienda Ospedaliero-Universitaria Senese, Italy
[f] Independent Medical Scientist, Milan, Italy
[g] Department of Experimental Medicine, Sapienza University of Rome, Rome, Italy
[h] Université Côte d'Azur, Inria, CNRS, I3S, Maasai
[i] Independent Data Scientist, Milan, Italy
[j] Good Clinical Practice Alliance-Europe (GCPA) and Strategic Initiative for Developing Capacity in Ethical Review-Europe (SIDCER), Leuven, Belgium
[k] Unit of Endocrinology and Prevention and Care of Diabetes, Center for Applied, Biomedical Research, Department of Medical and Surgical Sciences, University of Bologna, S. Orsola-Malpighi Hospital, Bologna, Italy
[l] Molecular Diagnostic Unit, Fundación Rioja Salud, Logroño, La Rioja, Spain
[m] Neurometabolic Diseases Laboratory, Bellvitge Biomedical Research Institute (IDIBELL), Barcelona, Spain
[n] CIBERER, Centro de Investigación Biomédica en Red de Enfermedades Raras, ISCIII, Melchor Fernández Almagro, 3, 28029 Madrid, Spain
[o] Spanish Covid HGE
[p] Immunology Division, Genetics Department, Hospital Universitari Vall d'Hebron, Vall d'Hebron Research Institute, Vall d'Hebron Barcelona Hospital Campus, Universitat Autònoma de Barcelona (UAB), Barcelona, Catalonia, Spain, EU
[q] Pediatric Infectious Diseases and Immunodeficiencies Unit, Hospital Universitari Vall d'Hebron, Vall d'Hebron Research Institute, Barcelona, Catalonia, Spain
[r] Institute of Medical and Molecular Genetics (INGEMM)-IdiPAZ, Hospital Universitario La Paz-UAM Paseo de La Castellana, 261, 28046 Madrid, Spain
[s] Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain
[t] Department of Medical, Surgical and Neurological Sciences, University of Siena, Italy
[u] Department of Medicine, Baylor College of Medicine, Houston TX, USA

ARTICLE INFO

Article History:
Received 1 December 2020
Revised 24 January 2021
Accepted 2 February 2021
Available online xxx

Keywords:
Androgen receptor gene

ABSTRACT

Background: While SARS-CoV-2 similarly infects men and women, COVID-19 outcome is less favorable in men. Variability in COVID-19 severity may be explained by differences in the host genome.
Methods: We compared poly-amino acids variability from WES data in severely affected COVID-19 patients versus SARS-CoV-2 PCR-positive oligo-asymptomatic subjects.
Findings: Shorter polyQ alleles (≤22) in the androgen receptor (*AR*) conferred protection against severe outcome in COVID-19 in the first tested cohort (both males and females) of 638 Italian subjects. The association between long polyQ alleles (≥23) and severe clinical outcome ($p = 0.024$) was also validated in an independent cohort of Spanish men <60 years of age ($p = 0.014$). Testosterone was higher in subjects with *AR* long-

* Corresponding author.
E-mail address: alessandra.renieri@unisi.it (A. Renieri).
[1] Co-first authors
[2] Co-last authors

https://doi.org/10.1016/j.ebiom.2021.103246

Testosterone
COVID-19
LASSO logistic regression
WES
Viral infection and host genome

polyQ, possibly indicating receptor resistance ($p = 0.042$ Mann-Whitney U test). Inappropriately low serum testosterone level among carriers of the long-polyQ alleles ($p = 0.0004$ Mann-Whitney U test) predicted the need for intensive care in COVID-19 infected men. In agreement with the known anti-inflammatory action of testosterone, patients with long-polyQ and age ≥60 years had increased levels of CRP ($p = 0.018$, not accounting for multiple testing).

Interpretation: We identify the first genetic polymorphism that appears to predispose some men to develop more severe disease. Failure of the endocrine feedback to overcome AR signaling defects by increasing testosterone levels during the infection leads to the polyQ tract becoming dominant to serum testosterone levels for the clinical outcome. These results may contribute to designing reliable clinical and public health measures and provide a rationale to test testosterone as adjuvant therapy in men with COVID-19 expressing long *AR* polyQ repeats.

Funding: MIUR project "Dipartimenti di Eccellenza 2018-2020" to Department of Medical Biotechnologies University of Siena, Italy (Italian D.L. n.18 March 17, 2020) and "Bando Ricerca COVID-19 Toscana" project to Azienda Ospedaliero-Universitaria Senese. Private donors for COVID-19 research and charity funds from Intesa San Paolo.

Research in context

Evidence before this study

We searched on Medline, EMBASE, and Pubmed for articles published from January 2020 to August 2020 using various combinations of the search terms "sex-difference", "gender" AND SARS-Cov-2, or COVID. Epidemiological studies indicate that men and women are similarly infected by COVID-19, but the outcome is less favorable in men, independently of age. Several studies also showed that patients with hypogonadism tend to be more severely affected. A prompt intervention directed toward the most fragile subjects with SARS-Cov-2 infection is currently the only strategy to reduce mortality. Glucocorticoid treatment is a cost-effective measure to improve the outcome of severe cases. Clinical algorithms have been proposed, but little is known on the ability of genetic profiling to predict outcome and disclose novel therapeutic strategies.

Added-value of this study

In a cohort of 1178 men and women with COVID-19, we used a supervised Machine Learning approach on a synthetic representation of genetic variability due to poly-amino acid repeats. Comparing the genotype of patients with extreme manifestations (severe vs. asymptomatic), we found an association between the poly-glutamine repeat number of the androgen receptor (AR) gene, serum testosterone concentrations, and COVID-19 outcome in male patients. Failure of the endocrine feedback to overcome AR signaling defects by increasing testosterone levels during the infection leads to the fact that polyQ ≥ 23 becomes dominant to testosterone levels for the clinical outcome.

Implications of all the available evidence

We identify the first genetic polymorphism predisposing some men to develop a more severe disease irrespectively of age. Based on this, we suggest that sizing the *AR* poly-glutamine repeat has important implications in the diagnostic pipeline of patients affected by life-threatening COVID-19 infection. Most importantly, our studies open to the potential of using testosterone as adjuvant therapy for patients with severe COVID-19 having defective androgen signaling, defined by this study as ≥23 PolyQ repeats, and inappropriately low levels of circulating androgens.

1. Introduction

Alongside the mode of transmission, viral load, comorbidities, and demographic factors (such as age and sex), the host genetic background appears to play an important role in COVID-19 severity and progression [1–8]. We hypothesized that common polymorphisms may contribute to COVID-19 severity, including poly-amino acids repeat polymorphisms, such as the polyQ tract of the Androgen Receptor (*AR*). AR contains in its N-terminus domain a polymorphic polyQ tract, ranging between 9 and 36 repeated CAG units in the normal population [9]. *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 [9–12].

Several lines of evidence lead to 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 [13,14]. 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 [15]. On the other hand, hypogonadism is known to correlate with severe COVID-19 [16] and other chronic conditions, partly due to the loss of attenuation of the inflammatory immune response exerted by testosterone (T) [17–19].

2. Methods

2.1. Patients

We performed a nested case-control study (NCC). Cases and controls were drawn from the Italian GEN-COVID cohort of 1178 subjects infected with SARS-CoV-2 diagnosed by RT-PCR on nasopharyngeal swab [2]. Demographic characteristics of patients enrolled in the cohort are summarized in Table 1 according to their clinical status. In the current NCC study, cases were selected according to the following inclusion criteria: *i.* CPAP/biPAP ventilation (230 subjects); *ii.* endotracheal intubation (108 subjects). As controls, 300 subjects were selected using the sole criterion of not requiring hospitalization. Exclusion criteria for both cases and controls were *i.* SARS-CoV-2 infection not confirmed by PCR; *ii.* non-caucasian ethnicity. Demographic characteristics of the subjects in the NCC study are summarized in Table 1. 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. All subjects were white European. The Spanish Covid HGE cohort is under IRB approval PR127/20 from Bellvitge University Hospital, Barcelona, Spain.

Table 1
Demographics characteristics of the Italian GEN-COVID Cohort and NCC study.

		Intubation	CPAP/BiPAP Ventilation	Oxygen Therapy	Hospitalized w/o respiratory support	Oligo^-asymptomatics w/o hospitalization
GEN-COVID	Number of Sybjects	108	230	352	188	300
	Male/Female	80/28	157/73	208/144	104/84	116/184
	Age males (years)	61,52±11,43	62,75±13,48	63,41±14,53	55,99±15,44	47,40±13,23
	Age females (years)	63,71±13,96	66,23±15,25	68,40±14,74	52,88±16,39	48,61±11,06
		Cases				**Controls**
NCC study	Number of Subjects	338				300
	Male/Female	237/101				116/184
	Age males (years)	62,34±12,84				47,40±13,23
	Age females (years)	65,53±14,94				48,61±11,06

^ Oligosymtpomatic: individuals with minor symptoms of COVID-19 (mild fever, cough, sore throat, etc.)

2.2. Ethics

The GEN-COVID study was approved by the University Hospital of Siena Ethics Review Board (Protocol n. 16917, dated March 16, 2020). This observational study has been inserted in www.clinicaltrial.org (NCT04549831). The Spanish Covid HGE cohort is under IRB approval PR127/20 from Bellvitge University Hospital, Barcelona Spain. Written informed consent was obtained from all individuals who contributed samples and data.

2.3. Analysis of triplets size in the AR locus

To establish allele sizes of the polymorphic triplet in the AR locus, we used the HUMARA assay with minor modifications [20]. Specifically, we performed a fluorescent PCR followed by capillary electrophoresis on an ABI3130 sequencer. Allele size was established using the Genescan Analysis software.

2.4. Binary representation of WES data

Variants calling was performed according to the GATK4 best practice guidelines, using BWA for mapping, and ANNOVAR for annotating. WES data were represented in a binary mode on a gene-by-gene basis. Poly-amino acids triplet repeats were represented in a binary mode: long and short repeats in respect to the reference sequence on the genome. A total of 40 genes with 43 triplet repeat regions were taken from UniProtKB (**Supplementary Table S1**). In the boolean representation of poly-amino acids triplet repeats, for each of these 40 genes two features were defined, Dij and Iij, with Dij being equal to 1 if gene i in sample j has a repeated region shorter than the reference, 0 otherwise, and Iij being equal to 1 if gene i in sample j has a repeated region longer than the reference, 0 otherwise.

2.5. LASSO logistic regression

We adopted the LASSO logistic regression that provides a feature selection method within the classification tasks able to enforce both the sparsity and the interpretability of the results. The weights of the logistic regression algorithm can be interpreted as the importance of the subset of the most relevant features for the task [21].

The input features of the LASSO logistic regression are the poly-amino acids triplet repeats as well as gender, comorbidity (1 if there is at least one comorbidity) and age, the latter as a continuous variable normalized between 0 and 1. Comorbidities were defined as the presence of one or more clinical conditions (i.e. cardiac, endocrine, neurological, neoplastic diseases) at the time of infection. During the fitting procedure, the class slight unbalancing is tackled by penalizing the misclassification of the minority class with a multiplicative factor inversely proportional to the class frequencies. The data pre-processing was coded in Python, whereas for the logistic regression model we used the scikit-learn module with the liblinear coordinate descent optimization algorithm.

2.6. Total T measurement

Blood samples were collected after an overnight fast, immediately centrifuged at 4 °C and stored at -20 °C until assayed. Serum and plasma total T (TT), SHBG levels in plasma and serum LH were measured following standard procedures.

Serum TT was measured using the Access testosterone assay (Beckman Coulter Inc., Fullerton, CA, USA) with a minimum detection limit of 0.35 nmol/L. Reference range for this assay was 6.07−27.1 nmol/L and liquid chromatography - tandem mass spectrometry (LC-MS/MS) according to a previously validated method provided with reference values between 9.8−28.4 nmol/L [22]. Thawed plasma underwent 15 min incubation at 56 °C for virus inactivation, and TT measured in 100 μl of plasma, with sensitivity limit being 0.270 nmol/L, imprecision ranging 9.8 to 0.7% and accuracy 90.6 to 101.5% at concentration levels between 1.12 and 39.2 nmol/L. A stability test under viral inactivation conditions was performed in 6 samples, revealing a T mean (min-max) % loss of 9.7% (4.6−16.7%).

SHBG levels were measured in plasma samples using Quantikine ELISA Kit (DSHB G0B, R&D Systems, Minneapolis, MN, USA) according to the manufacturers' instructions. Serum LH was measured using "Access LH assay" a chemiluminescenSert, two-step enzyme immunoassay (Beckman Coulter Inc., Fullerton, CA, USA). Sensitivity for the LH determination is 0.2 mIU/mL. Reference range in adult males for this assay is 1.2−8.6 mIU/mL.

2.7. Statistical analysis

Since serum and plasma T values were not normally distributed, the statistical analyses were performed using non-parametric tests. When appropriate, transformation was used for skewed data in regression models. We used the Mann-Whitney U test to compare T levels in males with *AR* long-polyQ (≥23) versus males with short-polyQ repeat (≤22). Logistic regression analysis was performed to test the contribution of age, T, and the number of polyglutamine repetitions on COVID-19 outcome. The only prespecified interaction tested was the T by polyQ (categorical). Box-Tidwell procedure was used to assess linearity and the Hosmer and Lemeshow to assess goodness of fit test. Multicollinearity was assessed by variance inflation factor, and dealt with by dropping the offending variables from the analysis on the basis of clinical grounds.

2.8. Role of funders

The work was financially supported by MIUR project "Dipartimenti di Eccellenza 2018-2020" to Department of Medical Biotechnologies University of Siena, Italy (Italian D.L. n.18 March 17, 2020) and by "Bando Ricerca COVID-19 Toscana" project to Azienda Ospedaliero-Universitaria Senese. It was also funded by private donors for COVID-19 research and charity funds from Intesa San Paolo "Fondo di Beneficenza n. b/2020/0119". The sponsors of the study had no role in study design, data collection, data analysis, data interpretation, or

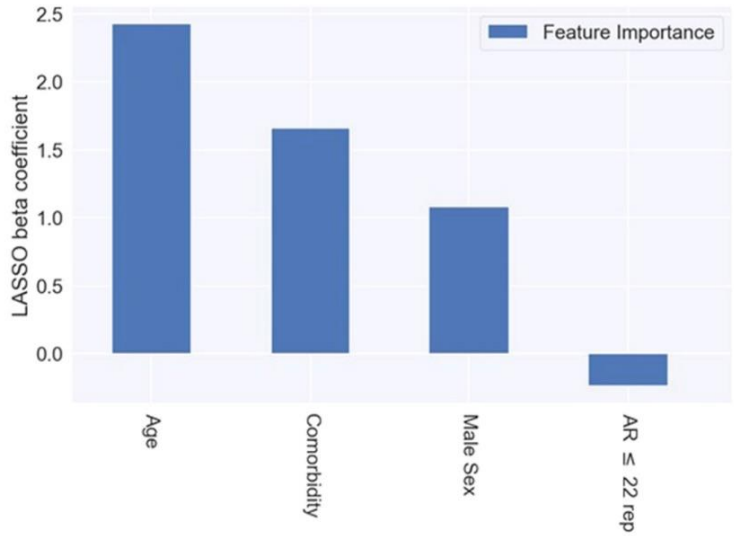
writing of the manuscript. The authors collected the data, and had full access to all of the data in the study. They also had the final decision and responsibility to submit the study results for publication.

3. Results

3.1. Testing the role of common poly-amino acid repeat polymorphisms in COVID-19 outcome

In order to test the role of common poly-amino acid repeat polymorphisms in determining COVID-19 clinical severity, we performed a NCC, selecting the extreme phenotypic ends of our entire GEN-COVID cohort (Table 1 and Fig. 1). Among 18,439 annotated genes, we selected those with amino acid repeats, namely 40 genes, and represented them as a boolean variable. Logistic regression with LASSO regularization analysis identified *AR* as the only protective gene (Fig. 1, panel a). The 10-fold cross-validation provides good performances in terms of accuracy (77%), precision (81%), sensitivity (77%), specificity (78%) and Area Under the Curve (AUC) score (86%) (Fig. 1, panel b). The performances of the logistic regression without LASSO regularization for the selected set of features (age, gender, comorbidity and AR gene) are 79% accuracy, 81% precision, 81% sensitivity, 78% specificity, 88% roc-auc. The model shows a slight decrease of almost all the performance measures when the AR gene is removed from the set (accuracy -1.2%, precision -1.3%, sensitivity -1.4%, specificity -1.2%, roc-auc +0.3%). Finally, the logistic regression

A)



B)

Performance	Average	Standard Deviation
Accuracy	77%	6%
Precision	81%	7%
Sensitivity	77%	7%
Specificity	78%	10%
Roc-auc	86%	6%

Fig. 1. LASSO logistic regression. The bar of the LASSO logistic regression beta coefficients represents the importance of each feature for the classification task (Fig. 1) (**Panel a**). The positive beta coefficients of the LASSO (upward bars) reflect a susceptible behaviour of the features to the target COVID-19 disease, whereas the negative coefficients (downward bars) a protective action. The calculated odd ratio of AR short repeats (≤22) is 0.79 i.e. protective. Therefore, the odd ratio of long repeats (≥23) is 1/0.79 = 1.27 i.e. severity. **Panel b:** Table reporting the averages and the standard deviations of accuracy, precision, sensitivity, specificity, and ROC-AUC scores for the 10-folds of the cross-validation.

on the male cohort with the AR gene alone provides results quite higher than the random guess (accuracy 58%, precision 71%, sensitivity 64%, specificity 55%, roc-auc 55%).

3.2. Validation of polyQ polymorphism by sizing the PolyQ repeat of the AR gene

In order to validate the results on *AR* obtained by LASSO logistic regression, we sized the number of triplets 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) [20]. We identified a 98% concordance between the results of the two techniques in measuring the polyQ repeats. 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 [23-24]. We 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 χ^2 test) (Table 2; **Supplementary Table S2**).

3.3. Validation of polyQ polymorphism in the Spanish Cohort

We then sized the polyQ repeat in an independent cohort consisting of 158 <60 years old Spanish males without known comorbidities (117 cases and 41 controls). The association with shorter repeats (≤ 22) and protection was confirmed (p-value 0.014 by χ^2 test) (Table 3).

3.4. Males with longer polyQ have receptor resistance

To functionally link the length of the PolyQ repeats to *AR* functionality, we measured TT in 183 men using LCMS/MS (**Supplementary Table S2**). TT was higher in patients carrying ≥23 vs ≤22 glutamines (13.45 vs 11.23 nmol/L, *p*-value 0.042), reflecting reduced negative feedback from the less active receptors present in patients carrying a PolyQ repeat of ≥23. This difference was evident also comparing the TT value and polyQ repeats in the case and the control group (Fig. 2).

3.5. Unbalanced T-AR axis in males with longer polyQ repeats

The hormonal status of the entire male cohort revealed lower TT and calculated free T levels and higher SHBG levels with increasing age (**Supplementary Table S3**).

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® [25]. Interviews were available for 61 subjects (43 short and 18 long) representative of the extremes genotypes (≤19 and ≥25 repeats) of the cohort. An Androtest score ≥8 was found in 38% of men with longer repeats as compared to 16% of those with ≤19 glutamines (likelihood ratio, $p = 0.046$). Similarly, cryptorchidism (11% in long repeats vs. 2% in short repeats), and anemia (11% in long repeats vs. 2% in short repeats), two powerful sings of low androgenicity, and severe erectile dysfunction (22% in long repeats vs. 9% in short repeats) were more frequently reported in subjects with longer repeats, but not osteopenia/osteoporosis (6% in long repeats vs. 7% in short repeats) (**Supplementary Table S4**). These results indicate a trend toward clinical hypogonadism for those with longer repeats. Conversely, in the entire male dataset, 6 cases of prostate cancer were found annotated in the past-medical history, all in the ≤22 glutamines group, suggesting an increased prostate sensitivity to androgens in this group. No difference was found in the prevalence of BPH or 5-alpha-reductase inhibitors use.

As the reduced signal transduction of AR might be partially compensated by higher T levels, we tested whether the decreased AR negative feedback was sufficient to overcome larger polyQ repeats size (**Fig. 2**). Logistic regression was performed to investigate the joint effect of T level and polyglutamine receptor length on the likelihood that subjects require intensive care during COVID infections, adjusting for age in the model. The logistic regression model was highly significant (χ^2 (3) = 18,881, $p < 0.0001$), with the model explaining 7.5% (Nagelkerke /R2) of the variance in COVID-19 outcome (**Supplementary Table S5**). To test whether the association between T and the outcome changes when the polyQ is short (≤ 22) or long (≥23), an interaction term was included in the model. A significant interaction was found (*p*-value 0.018), suggesting impaired feedback as a predictor of the worst outcome, namely intubation or CPAP/BiPAP versus hospitalization not requiring respiratory assistance. To provide an intuitive graphical representation, we plotted the ratio between TT serum concentrations and polyQ number vs. clinical outcome (**Supplementary Figure 1**). Results show a decreased mean ratio, a sign of an inappropriate rise of TT for increasing polyglutamine repeats, and association with a worse outcome ($p = 0.0004$).

3.6. Inflammatory phenotype in males with longer polyQ repeats

Finally, we tested the relationship between the *AR* polyQ repeat size and 5 laboratory markers of immunity/inflammation, including CRP, Fibrinogen, IL6, CD4 and NK count. We found that older (≥60) males with *AR* polyQ tract ≥23 have a higher (55.92 versus 48.21 mg/dl) mean value of CRP (p-value 0.018, not accounting for multiple testing) and lower mean value of Fibrinogen and a trend of higher IL6 (Table 4).

4. Discussion

We employed machine learning methodologies to identify a set of genes involved in the severity of COVID-19. In the presence of very high dimensionality, as for instance in a WES study, it is crucial to select the most predictive genes representing patterns of variation (mutations or variants) in subjects with different classes of response (i.e., disease state: from asymptomatic to severe cases). This problem is even more complex in diseases where multiple genes are involved in determining the severity and clinical variability of the pathology. Here, we wanted to represent poly-amino acids repeat

Table 2
PolyQ alleles correlation with COVID-19 outcome - males with age <60.

Males <60	<22	>23	*Marginal Row Totals*
Cases	52 (59,1%)	36 (40,9%)	88 (48,1%)
Controls	71 (74,7%)*	24 (25,3%)	95 (51,9%)
Marginal Column Totals	123 (67,2%)	60 (32,8%)	183 (Grand Total)

* p-value (cases vs controls) =0.024

Table 3
Validation in Spanish cohort

Spanish validation (χ2) Males global	≤22	≥23	*Marginal Row Totals*
Cases	51 (43,6%)	66 (56,4%)	117 (74,1%)
Controls	27 (65,9%)*	14 (34,1%)	41 (25,9%)
Marginal Column Totals	78 (49,4%)	80 (50,6%)	158 (Grand Total)

* **p-value (cases vs controls)=**0.014 (Significant at $p<0.05$)

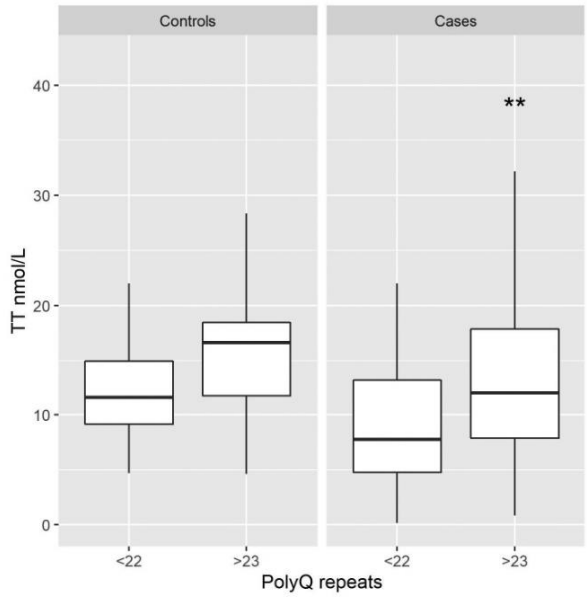


Fig. 2. Relationship between Total Testosterone and polyQ repeats in the case and the control group. Box-plot showing values of Total Testosterone (TT), expressed in nmol/L, in subjects with shorter (≤22) and longer (≥23) polyQ repeats in AR gene grouped between controls (left panel) and cases (right 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).

polymorphisms that are typically missed in classical GWAS analysis, which concentrates on bi-allelic polymorphisms.

We used a machine learning approach and logistic regression with a LASSO regularization to test if using such a simplified representation could lead to a reliable prediction of extreme clinical outcomes (asymptomatic versus severely affected). This approach enabled us to predict such clinical outcomes with 77% sensitivity.

AR contains a highly variable polyglutamine repeat (poly-Q) located in the N-terminal domain of the protein, spanning from 9 to 36 glutamine residues in the normal population [5]. *AR* polyQ length correlates with receptor functionality, with shorter polymorphic glutamine repeats typically associated with higher and longer PolyQ tracts with lower receptor activity [5]. *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 we decided to focus on male patients and demonstrated that shorter polymorphic glutamine repeats (≤22) confer protection against life-threatening COVID-19 in a subpopulation of individuals with age <60 years.

We also confirmed the association between polyQ size and receptor activity. Specifically, we showed that longer polyQ size (≥23) is associated with higher serum T levels, suggestive of impaired negative feedback (p=0.004 at Mann-Whitney U test) at the level of the hypothalamus and pituitary gland. While this is compensated in healthy subjects [26], during non-gonadal illnesses (NGI) such as COVID-19, some patients are unable to compensate for the reduced AR activity with higher T levels [27]. The result is a status of reduced androgenicity even in the presence of apparently normal T values [27].

As T is known to have an immunomodulatory activity attenuating inflammatory immune responses [26–32], we hypothesized that a long PolyQ repeat would lead to a pro-inflammatory status heralded by increased proinflammatory markers [19,33] 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. We found that -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.

The allele distribution of the PolyQ repeat length varies among different populations, with the shortest in Africans, medium in Caucasians, and longest in Asians [34]. Interestingly, WHO data on mortality rates during the first pandemic wave indicated a higher fatality rate in China and Italy (https://covid19.who.int/) [35] with respect to African. Hence, *AR* polyQ length variability could represent an

Table 4
Correlation between polyQ repeats in AR gene and laboratory values

CRP M≥60y cases			CRP M<60y cases		
Triplets	Mean	Count	Triplets	Mean	Count
≤**22**	48,21	78	≤**22**	54,5	43
≥**23**	55,92	38	≥**23**	26,41	29
p-value = 0.018 (Significant at p<0,05)			p-value = 0.2		

Fibrinogen M≥60y cases			Fibrinogen M<60y cases		
Triplets	Mean	Count	Triplets	Mean	Count
≤**22**	401,33	57	≤**22**	316,93	22
≥**23**	320,34	27	≥**23**	356,91	19
p-value = 0.093			*p*-value = 0.53		

IL6 times the upper limit of normal M≥60y cases			IL6 times the upper limit of normal M<60y cases		
Triplets	Mean	Count	Triplets	Mean	Count
≤**22**	54,56	40	≤**22**	40,43	17
≥**23**	75,78	16	≥**23**	31,8	14
p-value = 0,249			*p*-value = 0,81		

CD4 Lymphocytes M≥60y cases			CD4 Lymphocytes M<60 cases		
Triplets	Mean	Count	Triplets	Mean	Count
≤**22**	264,06	32	≤**22**	503,68	16
≥**23**	357,52	21	≥**23**	396,13	15
p-value = 0.22			*p*-value = 0.45		

NK Cells M≥60y cases			NK Cells M<60y cases		
Triplets	Mean	Count	Triplets	Mean	Count
≤**22**	70,71	28	≤**22**	147,3	13
≥**23**	102,25	16	≥**23**	107,14	14
p-value = 0.179			*p*-value = 0.098		

explanation for the observed differences in death rate. Moreover, Africans seem to be more prone to infection [36]. This observation could be due to a more active AR receptor, leading to a higher expression of *TMPRSS2*, a protease essential for SARS-CoV-2 spread [15].

Different studies have shown an association between hypogonadism or long polyQ repeats and severe COVID-19 [16,37] and other chronic obstructive pulmonary diseases [17,18]. Our results are in line with these initial observations and provide a possible mechanism explaining these associations. The present study brings these observations to the next level, revealing that is the overall androgenic effect -resulting from the interaction of polyQ polymorphism and circulating T levels- that predicts the need for intensive care. In infected men, we observed impaired feedback no longer sufficient to compensate for the reduced AR transcriptional activity, leading to the conclusion that polyQ tract length is hierarchically more important than serum T levels. This concept helps to solve some inconsistencies, including the early reports of a slightly better outcome in prostate cancer patients -who tend to have smaller polyQ repeats, as in our cohort - when compared to other cancers. Interestingly, previous studies failed to link polyQ with mortality, in healthy subjects [26] or individuals with chronic diseases such as diabetes mellitus [38]. Thus, the observed association between low androgenicity and outcome seems related to the hyperinflammatory state present in severe COVID-19.

An improvement in peak oxygen saturation in men receiving T replacement therapy has been demonstrated in a randomized controlled trial [39] and could be one of the mechanisms responsible for the observed protective effect of AR's with shorter polyQ tract in COVID-19 patients. The observations reported in this study prompt organizing a clinical trial where patients selected based on their serum T concentration and polyQ repeat size are randomized to receive T vs. placebo. Such study could introduce the concept that 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.

Variants of another X-linked gene, *TLR7*, have been associated with severe COVID-19 outcomes in young men [6]. In the 2 reported families, the rare *TLR7* mutations segregated as a highly penetrant monogenic X-linked recessive trait. While variants in *TLR7* gene are expected to account for a small number of severely affected cases, our findings involve a much larger number of subjects, as long polyQ alleles are relatively common (27%) [40]. Overall, X-linked genetic variants keep coming up as important for defining severe COVID-19 cases in males.

In conclusion, we present a method that can predict if subjects infected by SARS-CoV-2 are at risk for life-threatening complications. This approach has 77% accuracy, 81% precision, 77% sensitivity, and 78% specificity. Furthermore, we present evidence suggesting that a more active AR has the potential to confer protection against COVID-19 severity. If confirmed, these observations should be followed by properly conducted clinical trials exploring if T replacement may decrease morbidity and mortality in patients affected by the most severe forms of the disease. Finally, as shown by regression analysis, ORs ranges between 1.26 and 1.45, therefore the risk of carrying a longer *AR* is much smaller than other already known strong predictors such as age and sex, but still is highly significant, relatively common, and among the very few known genetic predictors of COVID-19 outcome.

Declaration of Competing Interest

The authors declare no competing interests.

Additional information

GEN-COVID Multicenter Study (https://sites.google.com/dbm.unisi.it/gen-covid)

Francesca Montagnani[2,22], Laura Di Sarno[1,2], Andrea Tommasi[1,2,5], Maria Palmieri[1,2], Massimiliano Fabbiani[22], Barbara Rossetti[22], Giacomo Zanelli[2,22], Fausta Sestini[20], Laura Bergantini[23], Miriana D'Alessandro[23], Paolo Cameli[23], David Bennet[23], Federico Anedda[24], Simona Marcantonio[24], Sabino Scolletta[24], Federico Franchi[24], Maria Antonietta Mazzei[25], Susanna Guerrini[25], Edoardo Conticini[26], Luca Cantarini[26], Bruno Frediani[26], Danilo Tacconi[27], Chiara Spertilli[27], Marco Feri[28], Alice Donati[28], Raffaele Scala[29], Luca Guidelli[29], Genni Spargi[30], Marta Corridi[30], Cesira Nencioni[31], Leonardo Croci[31], Gian Piero Caldarelli[32], Maurizio Spagnesi[33], Paolo Piacentini[33], Maria Bandini[33], Elena Desanctis[33], Silvia Cappelli[33], Anna Canaccini[34], Agnese Verzuri[34], Valentina Anemoli[34], Agostino Ognibene[35], Massimo Vaghi[36], Antonella D'Arminio Monforte[37], Esther Merlini[37], Federica Gaia Miraglia[37], Mario U. Mondelli[38,39], Stefania Mantovani[38], Serena Ludovisi[38,39], Massimo Girardis[40], Sophie Venturelli[40], Marco Sita[40], Andrea Cossarizza[41], Andrea Antinori[42], Alessandra Vergori[42], Arianna Emiliozzi[2,22,42], Stefano Rusconi[43,44], Matteo Siano[44], Arianna Gabrieli[44], Agostino Riva[43,44], Daniela Francisci[45,46], Elisabetta Schiaroli[45], Pier Giorgio Scotton[47], Francesca Andretta[47], Sandro Panese[48], Stefano Baratti[48] Renzo Scaggiante[49], Francesca Gatti[49], Saverio Giuseppe Parisi[50], Francesco Castelli[51], Eugenia Quiros-Roldan[51], Melania Degli Antoni[51], Isabella Zanella[52], Matteo Della Monica[53], Carmelo Piscopo[53], Mario Capasso[54,55,56], Roberta Russo[54,55], Immacolata Andolfo[54,55], Achille Iolascon[54,55], Giuseppe Fiorentino[57], Massimo Carella[58], Marco Castori[58], Giuseppe Merla[58], Filippo Aucella[59], Pamela Raggi[60], Carmen Marciano[60], Rita Perna[60], Matteo Bassetti[61,62], Antonio Di Biagio[62], Maurizio Sanguinetti[63,64], Luca Masucci[63,64], Serafina Valente[65], Maria Antonietta Mencarelli[5], Caterina Lo Rizzo[5], Elena Bargagli[23], Marco Mandalà[66], Alessia Giorli[66], Lorenzo Salerni[66], Patrizia Zucchi[67], Pierpaolo Parravicini[67], Elisabetta Menatti[68], Tullio Trotta[69], Ferdinando Giannattasio[69], Gabriella Coiro[69], Fabio Lena[70], Domenico A. Coviello[71], Cristina Mussini[72], Giancarlo Bosio[73], Enrico Martinelli[73], Sandro Mancarella[74], Luisa Tavecchia[74], Lia Crotti[75,76,77,78,79]

[22]Dept of Specialized and Internal Medicine, Tropical and Infectious Diseases Unit, Azienda Ospedaliera Universitaria Senese, Siena, Italy

[23]Unit of Respiratory Diseases and Lung Transplantation, Department of Internal and Specialist Medicine, University of Siena, Italy

[24]Dept of Emergency and Urgency, Medicine, Surgery and Neurosciences, Unit of Intensive Care Medicine, Siena University Hospital, Italy

[25]Department of Medical, Surgical and Neuro Sciences and Radiological Sciences, Unit of Diagnostic Imaging, University of Siena, Italy

[26]Rheumatology Unit, Department of Medicine, Surgery and Neurosciences, University of Siena, Policlinico Le Scotte, Italy

[27]Department of Specialized and Internal Medicine, Infectious Diseases Unit, San Donato Hospital Arezzo, Italy

[28]Dept of Emergency, Anesthesia Unit, San Donato Hospital, Arezzo, Italy

[29]Department of Specialized and Internal Medicine, Pneumology Unit and UTIP, San Donato Hospital, Arezzo, Italy

[30]Department of Emergency, Anesthesia Unit, Misericordia Hospital, Grosseto, Italy

[31]Department of Specialized and Internal Medicine, Infectious Diseases Unit, Misericordia Hospital, Grosseto, Italy

[32]Clinical Chemical Analysis Laboratory, Misericordia Hospital, Grosseto, Italy

[33]Department of Preventive Medicine, Azienda USL Toscana Sud Est, Italy

[34]Territorial Scientific Technician Department, Azienda USL Toscana Sud Est, Italy

[35]Clinical Chemical Analysis Laboratory, San Donato Hospital, Arezzo, Italy

[36]Chirurgia Vascolare, Ospedale Maggiore di Crema, Italy

[37]Department of Health Sciences, Clinic of Infectious Diseases, ASST Santi Paolo e Carlo, University of Milan, Italy

[38]Division of Infectious Diseases and Immunology, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy

[39]Department of Internal Medicine and Therapeutics, University of Pavia, Italy

[40]Department of Anesthesia and Intensive Care, University of Modena and Reggio Emilia, Modena, Italy

[41]Department of Medical and Surgical Sciences for Children and Adults, University of Modena and Reggio Emilia, Modena, Italy

[42]HIV/AIDS Department, National Institute for Infectious Diseases, IRCCS, Lazzaro Spallanzani, Rome, Italy

[43]III Infectious Diseases Unit, ASST-FBF-Sacco, Milan, Italy

[44]Department of Biomedical and Clinical Sciences Luigi Sacco, University of Milan, Milan, Italy

[45]Infectious Diseases Clinic, Department of Medicine 2, Azienda Ospedaliera di Perugia and University of Perugia, Santa Maria Hospital, Perugia, Italy

[46]Infectious Diseases Clinic, "Santa Maria" Hospital, University of Perugia, Perugia, Italy

[47]Department of Infectious Diseases, Treviso Hospital, Local Health Unit 2 Marca Trevigiana, Treviso, Italy

[48]Clinical Infectious Diseases, Mestre Hospital, Venezia, Italy.

[49]Infectious Diseases Clinic, ULSS1, Belluno, Italy

[50]Department of Molecular Medicine, University of Padova, Italy

[51]Department of Infectious and Tropical Diseases, University of Brescia and ASST Spedali Civili Hospital, Brescia, Italy

[52]Department of Molecular and Translational Medicine, University of Brescia, Italy; Clinical Chemistry Laboratory, Cytogenetics and Molecular Genetics Section, Diagnostic Department, ASST Spedali Civili di Brescia, Italy

[53]Medical Genetics and Laboratory of Medical Genetics Unit, A.O. R.N. "Antonio Cardarelli", Naples, Italy

[54]Department of Molecular Medicine and Medical Biotechnology, University of Naples Federico II, Naples, Italy

[55]CEINGE Biotecnologie Avanzate, Naples, Italy

[56]IRCCS SDN, Naples, Italy

[57]Unit of Respiratory Physiopathology, AORN dei Colli, Monaldi Hospital, Naples, Italy

[58]Division of Medical Genetics, Fondazione IRCCS Casa Sollievo della Sofferenza Hospital, San Giovanni Rotondo, Italy

[59]Department of Medical Sciences, Fondazione IRCCS Casa Sollievo della Sofferenza Hospital, San Giovanni Rotondo, Italy

[60]Clinical Trial Office, Fondazione IRCCS Casa Sollievo della Sofferenza Hospital, San Giovanni Rotondo, Italy

[61]Department of Health Sciences, University of Genova, Genova, Italy

[62]Infectious Diseases Clinic, Policlinico San Martino Hospital, IRCCS for Cancer Research Genova, Italy

[63]Microbiology, Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Catholic University of Medicine, Rome, Italy

[64]Department of Laboratory Sciences and Infectious Diseases, Fondazione Policlinico Universitario A. Gemelli IRCCS, Rome, Italy

[65]Department of Cardiovascular Diseases, University of Siena, Siena, Italy

[66]Otolaryngology Unit, University of Siena, Italy

[67]Department of Internal Medicine, ASST Valtellina e Alto Lario, Sondrio, Italy

[68]Study Coordinator Oncologia Medica e Ufficio Flussi Sondrio, Italy
[69]First Aid Department, Luigi Curto Hospital, Polla, Salerno, Italy
[70]Local Health Unit-Pharmaceutical Department of Grosseto, Toscana Sud Est Local Health Unit, Grosseto, Italy
[71]U.O.C. Laboratorio di Genetica Umana, IRCCS Istituto G. Gaslini, Genova, Italy.
[72]Infectious Diseases Clinics, University of Modena and Reggio Emilia, Modena, Italy.
[73]Department of Respiratory Diseases, Azienda Ospedaliera di Cremona, Cremona, Italy
[74]U.O.C. Medicina, ASST Nord Milano, Ospedale Bassini, Cinisello Balsamo (MI), Italy
[75]Istituto Auxologico Italiano, IRCCS, Department of Cardiovascular, Neural and Metabolic Sciences, San Luca Hospital, Milan, Italy.
[76]Department of Medicine and Surgery, University of Milano-Bicocca, Milan, Italy
[77]Istituto Auxologico Italiano, IRCCS, Center for Cardiac Arrhythmias of Genetic Origin, Milan, Italy.
[78]Istituto Auxologico Italiano, IRCCS, Laboratory of Cardiovascular Genetics, Milan, Italy
[79]Member of the European Reference Network for Rare, Low Prevalence and Complex Diseases of the Heart-ERN GUARD-Heart

Spanish COVID HGE

Sergio Aguilera-Albesa[80], Sergiu Albu[81], Carlos Casasnovas[82,13], Valentina Vélez-Santamaria[82,13], Juan Pablo Horcajada[83], Judit Villar[83], Agustí Rodríguez-Palmero[84,13,14], Montserrat Ruiz[13,14], Luis M Seijo[85], Jesús Troya[86], Juan Valencia-Ramos[87], Marta Gut[88]

[80]Navarra Health Service Hospital, Pamplona, Spain
[81]Institut Guttmann Foundation, Badalona, Barcelona, Spain
[82]Bellvitge University Hospital, L'Hospitalet de Llobregat, Barcelona, Spain
[83]Hospital del Mar, Parc de Salut Mar, Barcelona, Spain
[84]University Hospital Germans Trias i Pujol, Badalona, Barcelona, Spain
[85]Clínica Universitaria de Navarra, Madrid, Spain
[86]Infanta Leonor University Hospital, Madrid, Spain
[87]University Hospital of Burgos, Burgos, Spain
[88]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Carrer Baldiri i Reixac 4, 08028, Barcelona, Spain

Contributors

EF, FM, AR designed the study. CF and IM, were in charge of biological samples' collection and biobanking. MB, FF were in charge of clinical data collection. MB, FF, AR, and FM performed analysis/interpretation of clinical data. UP, FF and MM performed T measurement by LC-MS/MS. EDG, AS, AP and LPS performed the validation of association between shorter repeats and protection in a Spanish cohort. MM and AI critically reviewed the manuscript and interpreted clinical data/androgen physiopathological processes. SA and MB were in charge of DNA isolations from peripheral blood samples. FV, GD, AG, RT carried the sequencing experiments. EB, NP, SF, CG, MG and MS, performed bioinformatics and statistical analyses. EB, NP, SD, CF, and SC prepared Figures and Tables. EB, NP, AMP, FPC, AR, EF and FM wrote the manuscript. All authors have reviewed and approved the manuscript.

Acknowledgements

This study is part of the GEN-COVID Multicenter Study, https://sites.google.com/dbm.unisi.it/gen-covid, the Italian multicenter study aimed at identifying the COVID-19 host genetic bases. Specimens were provided by the COVID-19 Biobank of Siena, which is part of the Genetic Biobank of Siena, member of BBMRI-IT, of Telethon Network of Genetic Biobanks (project no. GTB18001), of EuroBioBank, and of RD-Connect. We thank the CINECA consortium for providing computational resources and the Network for Italian Genomes (NIG) http://www.nig.cineca.it for its support. We thank private donors for the support provided to A. R. (Department of Medical Biotechnologies, University of Siena) for the COVID-19 host genetics research project (D.L n.18 of March 17, 2020). We also thank the COVID-19 Host Genetics Initiative (https://www.covid19hg.org/), MIUR project "Dipartimenti di Eccellenza 2018-2020" to the Department of Medical Biotechnologies University of Siena, Italy and "Bando Ricerca COVID-19 Toscana" project to Azienda Ospedaliero-Universitaria Senese. We also thank Intesa San Paolo for the 2020 charity fund dedicated to the project N. B/2020/0119 "Identificazione delle basi genetiche determinanti la variabilità clinica della risposta a COVID-19 nella popolazione italiana".

Data availability and data sharing statement

The samples referenced here are housed in the GEN-COVID Patient Registry and the GEN-COVID Biobank and are available for sharing. The sequencing data are deposited in http://www.nig.cineca.it/, specifically, http://nigdb.cineca.it) and available for consultation. For further information, you may contact the corresponding author, Prof. Alessandra Renieri (e-mail: alessandra.renieri@unisi.it).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ebiom.2021.103246.

References

[1] Benetti E, Giliberti A, Emiliozzi A, et al. Clinical and molecular characterization of COVID-19 hospitalized patients. PLoS One 2020;15(11):e0242534 Published 2020 Nov 18. doi: 10.1371/journal.pone.0242534.
[2] Daga S, Fallerini C, Baldassarri M, et al. Employing a systematic approach to biobanking and analyzing clinical and genetic data for advancing COVID-19 research [published online ahead of print, 2021 Jan 17]. *Eur J Hum Genet*. 2021;1-15. doi:10.1038/s41431-020-00793-7.
[3] Zhang X, Tan Y, Ling Y, et al. Viral and host factors related to the clinical outcome of COVID-19. Nature 2020;583(7816):437–40. doi: 10.1038/s41586-020-2355-0.
[4] Ellinghaus D, Degenhardt F, Bujanda L, et al. Genomewide association study of severe Covid-19 with respiratory failure [published online ahead of print, 2020 Jun 17]. N Engl J Med. 2020 NEJMoa2020283. doi: 10.1056/NEJMoa2020283.
[5] Zhang Q, Bastard P, Liu Z, et al. Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. Science 2020:eabd4570. Epub ahead of print. PMID:32972995. doi: 10.1126/science.abd4570.
[6] van der Made CI, Simons A, Schuurs-Hoeijmakers J, et al. Presence of genetic variants among young men with severe COVID-19. JAMA 2020;324(7):1–11 Epub ahead of print. PMID:32706371; PMCID: PMC7382021. doi: 10.1001/jama.2020.13719.
[7] Bastard P, Rosen LB, Zhang Q, et al. Autoantibodies against type I IFNs in patients with life-threatening COVID-19. Science 2020;370(6515):eabd4585. Epub 2020 Sep 24. PMID:32972996. doi: 10.1126/science.abd4585.
[8] Pivonello R, Auriemma RS, Pivonello C, et al. Sex disparities in COVID-19 severity and outcome: are men weaker or women stronger? [published online ahead of print, 2020 Nov 26]. Neuroendocrinology 2020 10.1159/000513346. doi: 10.1159/000513346.
[9] Callewaert L, Christiaens V, Haelens A, Verrijdt G, Verhoeven G, Claessens F. Implications of a polyglutamine tract in the function of the human androgen receptor. Biochem Biophys Res Commun 2003;306(1):46–52. doi: 10.1016/s0006-291x(03)00902-1.
[10] Simanainen U, Brogley M, Gao YR, et al. Length of the human androgen receptor glutamine tract determines androgen sensitivity in vivo. Mol Cell Endocrinol 2011;342:81–6. doi: 10.1016/j.mce.2011.05.011.
[11] Tirabassi G, Cignarelli A, Perrini S, et al. Influence of CAG repeat polymorphism on the targets of testosterone action. Int J Endocrinol 2015;2015:298107. doi: 10.1155/2015/298107.
[12] Lindström S, Ma J, Altshuler D, et al. A large study of androgen receptor germline variants and their relation to sex hormone levels and prostate cancer risk. Results from the national cancer institute breast and prostate cancer cohort consortium. J Clin Endocrinol Metab 2010;95(9):E121–7. doi: 10.1210/jc.2009-1911.
[13] Wambier CG, Goren A. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection is likely to be androgen mediated. J Am Acad Dermatol 2020;83(1):308–9. doi: 10.1016/j.jaad.2020.04.032.
[14] Pozzilli P, Lenzi A. Commentary: testosterone, a key hormone in the context of COVID-19 pandemic. Metabolism 2020;108:154252. doi: 10.1016/j.metabol.2020.154252.

[15] Hoffmann M, Kleine-Weber H, Schroeder S, et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cell 2020;181(2):271–80 .e8. doi: 10.1016/j.cell.2020.02.052.

[16] Rastrelli G, Di Stasi V, Inglese F, et al. Low testosterone levels predict clinical adverse outcomes in SARS-CoV-2 pneumonia patients [published online ahead of print, 2020 May 20]. Andrology 2020 10.1111/andr.12821. doi: 10.1111/andr.12821.

[17] Van Vliet M, Spruit MA, Verleden G, et al. Hypogonadism, quadriceps weakness, and exercise intolerance in chronic obstructive pulmonary disease. Am J Respir Crit Care Med 2005;172(9):1105–11. doi: 10.1164/rccm.200501-114OC.

[18] Mohan SS, Knuiman MW, Divitini ML, et al. Higher serum testosterone and dihydrotestosterone, but not oestradiol, are independently associated with favourable indices of lung function in community-dwelling men. Clin Endocrinol (Oxf) 2015;83(2):268–76. doi: 10.1111/cen.12738.

[19] Mohamad NV, Wong SK, Wan Hasan WN, et al. The relationship between circulating testosterone and inflammatory cytokines in men. Aging Male 2019;22(2):129–40. doi: 10.1080/13685538.2018.1482487V.

[20] Allen RC, Zoghbi HY, Moseley AB, et al. Methylation of Hpall and Hhal sites near the polymorphic CAG repeat in the human androgen-receptor gene correlates with X chromosome inactivation. Am J Hum Genet. 1992;51(6):1229–39.

[21] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) 1996;58(21):267–88 .36.

[22] Fanelli F, Belluomo I, Di Lallo VD, et al. Serum steroid profiling by isotopic dilution-liquid chromatography-mass spectrometry: comparison with current immunoassays and reference intervals in healthy adults. Steroids 2011 Feb;76(3):244–53. doi: 10.1016/j.steroids.2010.11.005.

[23] Mengual L, Oriola J, Ascaso C, Ballescà JL, Oliva R. An increased CAG repeat length in the androgen receptor gene in azoospermic ICSI candidates. J Androl 2003;24(2):279–84. doi: 10.1002/j.1939-4640.2003.tb02673.x.

[24] Möhlig M, Arafat AM, Osterhoff MA, et al. Androgen receptor CAG repeat length polymorphism modifies the impact of testosterone on insulin sensitivity in men. Eur J Endocrinol 2011 Jun;164(6):1013–8. doi: 10.1530/EJE-10-1022.

[25] Millar AC, Lau ANC, Tomlinson G, Kraguljac A, Simel DL, Detsky AS, Lipscombe LL. Predicting low testosterone in aging men: a systematic review. CMAJ 2016 Sep 20;188(13):E321–30.

[26] Eendebak RJ, Huhtaniemi IT, Pye SR, et al. The androgen receptor gene CAG repeat in relation to 4-year changes in androgen-sensitive endpoints in community-dwelling older European men. Eur J Endocrinol 2016;175(6):583–93. doi: 10.1530/EJE-16-0447.

[27] Zitzmann M, Nieschlag E. The CAG repeat polymorphism within the androgen receptor gene and maleness. Int J Androl 2003;26(2):76–83. doi: 10.1046/j.1365-2605.2003.00393.x.

[28] Gubbels Bupp MR, Jorgensen TN. Androgen-induced immunosuppression. Front Immunol 2018;9:794. Published 2018 Apr 17. doi: 10.3389/fimmu.2018.00794.

[29] Hoebe K, Janssen E, Beutler B. The interface between innate and adaptive immunity. Nat Immunol 2004;5(10):971–4. doi: 10.1038/ni1004-971.

[30] Janeway Jr. CA, Medzhitov R. Innate immune recognition. Annu Rev Immunol 2002;20:197–216. doi: 10.1146/annurev.immunol.20.083001.084359.

[31] Lai JJ, Lai KP, Zeng W, Chuang KH, Altuwaijri S, Chang C. Androgen receptor influences on body defense system via modulation of innate and adaptive immune systems: lessons from conditional AR knockout mice. Am J Pathol 2012;181(5):1504–12. doi: 10.1016/j.ajpath.2012.07.008.

[32] Medzhitov R, Janeway Jr. C. Innate immunity. N Engl J Med 2000;343(5):338–44. doi: 10.1056/NEJM200008033430506.

[33] Pierotti S, Lolli F, Lauretta R, et al. Androgen modulation of pro- inflammatory and anti-inflammatory cytokines during preadipocyte differentiation. Horm Mol Biol Clin Investig 2010;4(1):483–8. doi: 10.1515/HMBCI.2010.059.

[34] Ackerman CM, Lowe LP, Lee H, et al. Ethnic variation in allele distribution of the androgen receptor (AR) (CAG)n repeat. J Androl 2012;33(2):210–5. doi: 10.2164/jandrol.111.013391.

[35] Porcheddu R, Serra C, Kelvin D, Kelvin N, Rubino S. Similarity in case fatality rates (CFR) of COVID-19/SARS-COV-2 in Italy and China. J Infect Dev Ctries 2020;14(2):125–8 Published 2020 Feb 29. doi: 10.3855/jidc.12600.

[36] de Lusignan S, Dorward J, Correa A, et al. Risk factors for SARS-CoV-2 among patients in the oxford royal college of general practitioners research and surveillance centre primary care network: a cross-sectional study. Lancet Infect Dis 2020;20(9):1034–42. doi: 10.1016/S1473- 3099(20)30371-6.

[37] McCoy J, Wambier CG, Herrera S, et al. Androgen receptor genetic variant predicts COVID-19 disease severity: a prospective longitudinal study of hospitalized COVID-19 male patients [published online ahead of print, 2020 Sep 25]. J Eur Acad Dermatol Venereol 2020 10.1111/jdv.16956. doi: 10.1111/jdv.16956.

[38] Wong PW, Lee HM, Lau ESH, et al. Interactive effects of testosterone and the androgen receptor CAG repeat length polymorphism on cardiovascular- renal events and mortality in men with diabetes. Diabetes Metab Res Rev 2019;35(1):e3081. doi: 10.1002/dmrr.3081.

[39] Caminiti G, Volterrani M, Iellamo F, et al. Effect of long-acting testosterone treatment on functional exercise capacity, skeletal muscle performance, insulin resistance, and baroreflex sensitivity in elderly patients with chronic heart failure a double-blind, placebo-controlled, randomized study. J Am Coll Cardiol 2009;54(10):919–27. doi: 10.1016/j.jacc.2009.04.078.

[40] Canale D, Caglieresi C, Moschini C, et al. Androgen receptor polymorphism (CAG repeats) and androgenicity [published correction appears in Clin Endocrinol (Oxf). 2005 Oct; **63**(4):482]. *Clin Endocrinol* (Oxf) 2005;63(3):356–61. doi: 10.1111/j.1365-2265.2005.02354.x.

6. Association of Toll-like receptor 7 variants with life-threatening COVID-19 disease in males: findings from a nested case-control study

Recently, loss-of-function variants in *TLR7* were identified in males with severe COVID-19 with a mean age of 26 years. As age and male sex are two major risk factors for developing life-threatening COVID-19 after infection, we investigated whether the two reported families represent the tip of the iceberg of a subset of young COVID-19 male patients.

In the previous chapter LASSO logistic regression model was used to identify a gene's common variants that is predictive for the severe or the mild COVID-19 phenotype. Here we report the analyses carried out by applying LASSO logistic regression method to the rare and ultra-rare genetic variants on the X chromosome (R_X and UR_X Boolean features described in chapter 2, section 2.5.1). This study, along with other published studies, shows that COVID-19 segregates like an X-linked recessive disorder environmentally conditioned by SARS-CoV-2. This type of inheritance contributes to disease susceptibility in up to 2% of severe COVID-19 [60].

eLife

SHORT REPORT

Association of Toll-like receptor 7 variants with life-threatening COVID-19 disease in males: findings from a nested case-control study

Chiara Fallerini[1,2†], Sergio Daga[1,2†], Stefania Mantovani[3†], Elisa Benetti[2], Nicola Picchiotti[4,5], Daniela Francisci[6,7], Francesco Paciosi[6,7], Elisabetta Schiaroli[6], Margherita Baldassarri[1,2], Francesca Fava[1,2,8], Maria Palmieri[1,2], Serena Ludovisi[3,9], Francesco Castelli[10], Eugenia Quiros-Roldan[10], Massimo Vaghi[11], Stefano Rusconi[12,13], Matteo Siano[12], Maria Bandini[14], Ottavia Spiga[5,15], Katia Capitani[1,16], Simone Furini[2], Francesca Mari[1,2,8], GEN-COVID Multicenter Study[1], Alessandra Renieri[1,2,8]*, Mario U Mondelli[3,9], Elisa Frullanti[1,2]

[1]Medical Genetics, University of Siena, Siena, Italy; [2]Med Biotech Hub and Competence Center, Department of Medical Biotechnologies, University of Siena, Siena, Italy; [3]Division of Infectious Diseases and Immunology, Department of Medical Sciences and Infectious Diseases, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy; [4]Department of Mathematics, University of Pavia, Pavia, Italy; [5]University of Siena, DIISM-SAILAB, Siena, Italy; [6]Infectious Diseases Clinic, Department of Medicine 2, Azienda Ospedaliera di Perugia and University of Perugia, Santa Maria Hospital, Perugia, Italy; [7]Infectious Diseases Clinic, "Santa Maria" Hospital, University of Perugia, Perugia, Italy; [8]Genetica Medica, Azienda Ospedaliero-Universitaria Senese, Siena, Italy; [9]Department of Internal Medicine and Therapeutics, University of Pavia, Pavia, Italy; [10]Department of Infectious and Tropical Diseases, University of Brescia and ASST Spedali Civili Hospital, Brescia, Italy; [11]Chirurgia Vascolare, Ospedale Maggiore di Crema, Crema, Italy; [12]Department of Biomedical and Clinical Sciences Luigi Sacco, University of Milan, Milan, Italy; [13]III Infectious Diseases Unit, ASST-FBF-Sacco, Milan, Italy; [14]Department of Preventive Medicine, Azienda USL Toscana Sud Est, Siena, Italy; [15]Department of Biotechnology, Chemistry and Pharmacy, University of Siena, Siena, Italy; [16]Molecular Mechanisms of Oncogenesis, ISPRO Core Research Laboratory (CRL), Firenze, Italy

***For correspondence:** alessandra.renieri@unisi.it

[†]These authors contributed equally to this work

Group author details: GEN-COVID Multicenter Study See page 10

Competing interests: The authors declare that no competing interests exist.

Funding: See page 13

Received: 16 February 2021
Accepted: 24 February 2021
Published: 02 March 2021

Reviewing editor: Frank L van de Veerdonk, University Medical Center, Netherlands

Abstract

Background: Recently, loss-of-function variants in TLR7 were identified in two families in which COVID-19 segregates like an X-linked recessive disorder environmentally conditioned by SARS-CoV-2. We investigated whether the two families represent the tip of the iceberg of a subset of COVID-19 male patients.

Methods: This is a nested case-control study in which we compared male participants with extreme phenotype selected from the Italian GEN-COVID cohort of SARS-CoV-2-infected participants (<60 y, 79 severe cases versus 77 control cases). We applied the LASSO Logistic Regression analysis, considering only rare variants on young male subsets with extreme phenotype, picking up TLR7 as the most important susceptibility gene.

Results: Overall, we found TLR7 deleterious variants in 2.1% of severely affected males and in none of the asymptomatic participants. The functional gene expression profile analysis demonstrated a reduction in TLR7-related gene expression in patients compared with controls demonstrating an impairment in type I and II IFN responses.
Conclusions: Young males with TLR7 loss-of-function variants and severe COVID-19 represent a subset of male patients contributing to disease susceptibility in up to 2% of severe COVID-19.
Funding: Funded by private donors for the Host Genetics Research Project, the Intesa San Paolo for 2020 charity fund, and the Host Genetics Initiative.
Clinical trial number: NCT04549831.

Introduction

Coronavirus disease 2019 (COVID-19), a potentially severe systemic disease caused by coronavirus SARS-CoV-2, is characterized by a highly heterogeneous phenotypic presentation, with the large majority of infected individuals experiencing only mild or no symptoms. However, severe cases can rapidly evolve toward a critical respiratory distress syndrome and multiple organ failure (***Wu and McGoogan, 2020***). COVID-19 still represents an enormous challenge for the world's healthcare systems almost 1 year after the first appearance in December 2019 in Wuhan, Huanan, Hubei Province of China. Although older age and the presence of cardiovascular or metabolic comorbidities have been identified as risk factors predisposing to severe disease (***Hägg et al., 2020***), these factors alone do not fully explain differences in severity (***Stokes et al., 2020***). Stokes EK et al. reported that male patients show more severe clinical manifestations than females with a statistically significant ($p<0.00001$) higher prevalence of hospitalizations (16% versus 12%), ICU admissions (3% versus 2%), and deaths (6% versus 5%) (***Stokes et al., 2020***). These results are in line with other reports indicating that gender may influence disease outcome (***Garg et al., 2020***; ***Goodman et al., 2020***).

These findings suggest a role of host predisposing genetic factors in the pathogenesis of the disease, which may be responsible for different clinical outcomes as a result of different antiviral defense mechanisms as well as specific receptor permissiveness to virus and immunogenicity.

Recent evidence suggests a fundamental role of interferon genes in modulating immunity to SARS-CoV-2; in particular, rare variants have recently been identified in the interferon type I pathway that are responsible for inborn errors of immunity in a small proportion of patients and auto-antibodies against type I interferon genes in up to 10% of severe COVID-19 cases (***Zhang et al., 2020***; ***Bastard et al., 2020***).

Toll-like receptors (TLRs) are crucial components in the initiation of innate immune responses to a variety of pathogens, causing the production of pro-inflammatory cytokines (TNF-α, IL-1, and IL-6) and type I and II Interferons (IFNs), that are responsible for innate antiviral responses. In particular, the innate immunity is very sensitive in detecting potential pathogens, activating downstream signaling to induce transcription factors in the nucleus, promoting synthesis and release of type I and type II IFNs in addition to a number of other proinflammatory cytokines, and leading to a severe cytokine release syndrome which may be associated with a fatal outcome. Interestingly, among the different TLRs, TLR7 recognizes several single-stranded RNA viruses including SARS-CoV-2 (***Poulas et al., 2020***). We previously showed that another RNA virus, hepatitis C virus (HCV), is able to inhibit CD4 T cell function via Toll-like receptor 7 (TLR7) (***Mele et al., 2017***). Recently, ***van der Made et al., 2020*** have reported two independent families in which COVID-19 segregates like an X-linked recessive monogenic disorder conditioned by SARS-CoV-2 as an environmental factor.

Here, we performed a nested case-control study within our prospectively recruited GEN-COVID cohort with the aim to determine whether the two families described by van der Made et al. represent an ultra-rare situation or the tip of the iceberg of a larger subset of young male patients.

Materials and methods

Patients and samples

A subset of 156 young (<60 years) male COVID-19 patients was selected from the Italian GEN-COVID cohort of 1,178 SARS-CoV-2-infected participants (https://sites.google.com/dbm.unisi.it/gen-covid) (***Daga et al., 2021***). The study (GEN-COVID) was consistent with Institutional guidelines and approved

by the University Hospital (Azienda Ospedaliero-Universitaria Senese) Ethical Review Board, Siena, Italy (Prot n. 16929, dated March 16, 2020). We performed a nested case-control study (STREGA reporting guideline was used to support reporting of this study). Cases were selected according to the following inclusion criteria: i. male gender; ii. young age (<60 years); iii endotracheal intubation or CPAP/biPAP ventilation (79 participants). As controls, 77 participants were selected using the sole criterion of being oligo-asymptomatic not requiring hospitalization. Cases and controls represented the extreme phenotypic presentations of the GEN-COVID cohort. Exclusion criteria for both cases and controls were: i. SARS-CoV-2 infection not confirmed by PCR; ii. non-white ethnicity. Materials and methods details are listed in the Online Repository. A similar cohort from the second wave, composed of 83 young male COVID-19 patients, was used to expand the cohort.

Statistical methods

We adopted the LASSO logistic regression, one of the most common Machine Learning algorithms for classification, that provides a feature selection method within the classification task able to enforce both the sparsity and the interpretability of the results (***Tibshirani, 1996***). In fact, the coefficients of the logistic regression model are directly related to the importance of the corresponding features, and LASSO regularization shrinks close to zero the coefficients of features that are not relevant in predicting the response, reducing overfitting and giving immediate interpretability of the model predictions in terms of few feature importance.

The principal components analysis (PCA) was applied prior to the LASSO logistic regression in order to remove samples that were clear outliers with respect to the first three principal components from the following analyses (deviating more than five standard deviations from the average).

A 10-fold cross-validation method was applied in order to test the performances. It provides the partition of the dataset into 10 batches, then nine batches are exploited for the training of the LASSO logistic regression and the remaining batch as a test, by repeating this procedure 10 times. The performance metrics are averaged on the 10 testing sets in order to avoid overfitting. The confusion matrix is built by summing up the predictions of the 10 testing folds. During the fitting procedure, the class unbalancing is tackled by penalizing the misclassification of the minority class with a multiplicative factor inversely proportional to the class frequencies.

In order to evaluate the significance of the association between *TLR7* variants and COVID severity, the Fisher's Exact Test was used.

For the quantitative PCR assay, the fold changes in mRNA expression level per gene were compared between the individual patients and controls using an unpaired t test on the log-transformed fold changes. p Values < 0.05 were considered statistically significant.

In vitro peripheral blood mononuclear cell (PBMC) experiments

Peripheral blood mononuclear cells (PBMC) were isolated by Ficoll-Hypaque (GE Healthcare Bio-Sciences AB) density gradient centrifugation as previously described (***Mantovani et al., 2019***). 5×10^5 PBMC from COVID-19 patients 6 months after recovery and six unaffected male and female controls were stimulated for 4 hr with the TLR7 agonist imiquimod at 5 µg/mL or cell culture medium. Total RNA extraction was performed with RNeasy Plus Mini kit and gDNA eliminator mini spin columns (QIAGEN, Hilden, Germany), following the manufacturer's instructions. First-strand cDNA was synthesized from total RNA using High-Capacity cDNA Reverse Transcription Kit following the manufacturer's instructions (Thermo Fisher Scientific, Waltham, Massachusetts, United States). The Advanced Universal SYBR Green Supermix (BioRad, Redmond, WA, United States) was used. All reactions were performed in triplicates using the CFX96 Real-Time machine detection system (BioRad, Redmond, WA, United States) and each sample was amplified in duplicate. The following primers were used:

TLR7	Fw Primer	5'-CATCAAGAGGCTGCAGATTAAA-3'
	Rv Primer	5'-GAAAAGATGTTGTTGGCCTCA-3'
IFN-γ	Fw Primer	5'-TGACCAGAGCATCCAAAAGA-3'
	Rv Primer	5'-CTCTTCGACCTCGAAACAGC-3'

Continued on next page

IRF7	Fw Primer	5′-CCATCTTCGACTTCAGAGTCTTC-3′
	Rv Primer	5′-TCTAGGTGCACTCGGCACAG-3′
ISG15	Fw Primer	5′-GACAAATGCGACGAACCTCT-3′
	Rv Primer	5′-GAACAGGTCGTCCTGCACAC-3′
IFN-a	Fw Primer	5′-GACTCCATCTTGGCTGTGA-3′
	Rv Primer	5′-TGATTTCTGCTCTGACAACCT-3′
HRPT1	Fw Primer	5′-TGACACTGGCAAAACAATGCA-3′
	Rv Primer	5′-GGTCCTTTTCACCAGCAAGCT-3′

A total of 2.5 × 10^5 PBMC from COVID-19 patients and healthy controls were maintained in RPMI-1640 supplemented with 10% of FCS, 1% antibiotic antimycotic solution, 1% L-glutamine and 1% Sodium Pyruvate (Sigma-Aldrich, St. Louis, MO, USA) and stimulated in vitro for 4 hr with Lipopolysaccharide (LPS) at 1 µg/ml or cell culture medium and the Protein Transport Inhibitor GolgiStop (BD Biosciences, San Diego, CA, USA). After washing, PBMC were stained for surface cell marker using mouse anti-CD14PerCP-Cy5.5 (BD Biosciences) and anti-CD3BV605 (BD Biosciences) monoclonal antibody (mAb). Cells were fixed with BD Cytofix/Cytoperm and permeabilized with the BD Perm/Wash buffer (BD Biosciences) according to the manufacturer's instructions, in the presence of anti-IL6BV421 (BD Biosciences) mAb. Ex-vivo TLR7 intracellular expression was evaluated in PBMC from patients and controls by flow cytometry. 2,5 × 10^5 PBMC were stained for surface markers using anti-CD19BV605, anti-CD14PerCP-Cy5.5 and anti-CD3BV421 (BD Biosciences) mAbs. Cells were fixed and permeabilized in the presence of anti-TLR7 Alexa Fluor 488 (R and D System, Minneapolis, MN, USA) mAb or isotype control as described above. After staining cells were washed, immediately fixed in CellFix solution (BD Biosciences) and analysed. Cell acquisition was performed on a 12-color FACSCelesta (BD Biosciences, San Diego, CA, USA) instrument. Data analysis was performed with the Kaluza 2.1 software (Beckman Coulter).

Protein stability prediction

The protein structure of Human Toll Like Receptor, UniProtKB ID Q9NYK1 [https://www.uniprot.org/uniprot/Q9NYK1], was obtained by homology modeling using Swiss Model tool (***Waterhouse et al., 2018***). The selected template protein with 97% of sequence identity was the Crystal structure of monkey TLR7 with PDB ID 5GMF [https://www.rcsb.org/structure/5GMF]. The two Val to Asp missense mutations were analysed by using different protein stability predictors like Polyphen-2 (***Adzhubei et al., 2010***), SIFT (***Ng and Henikoff, 2003***), and DynaMut (***Rodrigues et al., 2018***).

Transfection experiments of TLR7 variants

PCR based site-directed mutagenesis was performed in pUNO-hTLR7 plasmid (Invivogen), kindly provided by Ugo D'Oro (GSK Vaccines, Siena, Italy) (***Iavarone et al., 2011***), to generate specific plasmids for each *TLR7* variant, including those considered neutral (mutagenic primers available on request).

All point mutations except for p.Arg920Lys were confirmed by Sanger sequencing. HEK293 cells were maintained in DMEM supplemented with 10% FBS, 1% L-Glutamine and 1% penicillin/streptomycin at 37°C with 5% CO_2. Transient transfections were performed using Lipofectamine 2000 (Invitrogen) according to manufacturer's instructions: 3 × 10^5 cell/well were seeded the day before, and then transfected with 2 µg of DNA. After 24 hr, the cells were stimulated with Imiquimod at 1 µg/ml for 4 hr and then total RNA was extracted with RNeasy Mini Kit (QIAGEN, Hilden, Germany). For each sample, cDNA was synthesized from 1 µg of total RNA using QantiTect Reverse Transcription kit (QIAGEN, Hilden, Germany) according to manufacturer's instructions. The expression of IFN-a in stimulated and unstimulated cells was evaluated by qRT-PCR using the same procedure as described for PBMCs.

Results and discussion

We applied 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 (https://sites.google.com/dbm.

eLife Short report Genetics and Genomics | Medicine

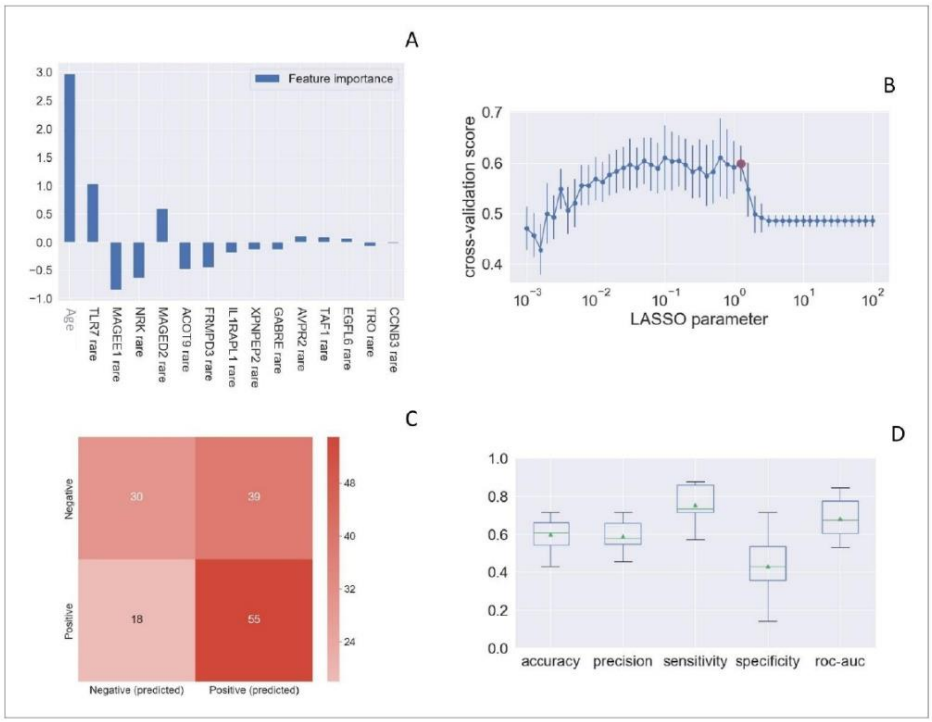


Figure 1. Rare *TLR7* variants and association with COVID-19. LASSO logistic regression on boolean representation of rare variants of all genes of the X chromosome is presented. *TLR7* is picked up by LASSO logistic regression as one of the most important genes on the X chr (Panel **A**). The LASSO logistic regression model provides an embedded feature selection method within the binary classification tasks (male patients with life-threatening COVID-19 vs infected asymptomatic male participants). The upward histograms (positive weights) reflect a susceptible behavior of the features to the target COVID-19, whereas the downward histograms (negative weights) a protective action. Panel **B** represents the cross-validation accuracy score for the grid of LASSO regularization parameters; the error bar is given by the standard deviation of the score within the 10 folds; the red circle (1.26) corresponds to the parameter chosen for the fitting procedure. Performances are evaluated through the confusion matrix of the aggregated predictions in the 10 folds of the cross-validation (Panel **C**) and with the boxplot (Panel **D**) of accuracy (60% average value), precision (59%), sensitivity (75%), specificity (43%), and ROC-AUC score (68%). The box extends from the Q1 to Q3 quartile, with a line at the median (Q2) and a triangle for the average.

Table 1. Fisher exact test of the overall combined cohorts in young males (<60 years).

Clinical category	N. wild-type variants (97.84%)	N. pathological variants (2.15%)	Total
Severely affected males	129	6	135
Asymptomatic males	104	0	104
Total	233	6	239 (Grand Total)

p-value=0.0037.

Table 2. *TLR7* variants in severely affected Italian males -all ages- (cases).

Nucleotide change	Amino acid change	dbSNP	CADD	ExAC_ NFE	Function*	N. of patients	Clinical category†	Age	Cohort	Patient ID
c.901T>C	Ser301Pro	-	26.4	N/A	LOF	1	3	46	Italian	P3
c.2759G>A	Arg920Lys	rs189681811	16.52	0.0002	LOF‡	1	4	49	Italian	P6
c.3094G>A	Ala1032Thr	rs147244662	22.3	0.0006	LOF	2	3	65/66	Italian	P7/P8
c.655G>A	Val219Ile	rs149314023	12.28	0.0003	HYPO	1	4	32	Italian	P1
c.863C>T	Ala288Val	rs200146658	15.37	0.000012	Neutral	1	3	57	Italian	P2
c.1343C>T	Ala448Val	rs5743781	13.08	0.00465	Neutral	2	3	53/58	Italian	P4/P5

CADD, Combined Annotation Dependent Depletion; ExAC, Exome Aggregation Consortium; NFE, Non-Finnish European;
*Function: HYPO, hypomorphic; LOF, loss-of-function;
†Clinical category: 4, Hospitalized and intubated; 3, Hospitalized and CPAP-BiPAP and high-flows oxygen treated; 2, Hospitalized and treated with conventional oxygen support only; 1, Hospitalized without respiratory support; 0, Not hospitalized oligo/asymptomatic individuals.
‡based on in silico prediction.

unisi.it/gen-covid) (***Daga et al., 2021***). The GEN-COVID study was consistent with Institutional guidelines and approved by the University Hospital (Azienda Ospedaliero-Universitaria Senese) Ethical Review Board, Siena, Italy (Prot n. 16929, dated March 16, 2020). Only rare variants (≤1% in European Non-Finnish population) were considered in the boolean representation: the gene was set to one if it included at least a missense, splicing, or loss-of-function rare variant, and 0 otherwise. Fisher Exact test was then used for the specific data validation.

Toll-like receptor 7 (*TLR7*) was picked up as one of the most important susceptibility genes by LASSO Logistic Regression analysis (***Figure 1***). We then queried the COVID-19 section of the Network of Italian Genome (NIG) database (http://www.nig.cineca.it/, specifically, http://nigdb.cineca.it) that houses the entire GEN-COVID cohort represented by more than 1000 WES data of COVID-19 patients and SARS-CoV-2 infected asymptomatic participants (***Bastard et al., 2020***). By selecting for young (<60 year-old) males, we obtained rare (MAF ≤ 1%) *TLR7* missense variants predicted to impact 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.

We then investigated a similar cohort coming from the Italian second wave composed of male patients under 60 years of age without comorbidities (56 cases and 27 controls) was used to expand the cohort. All participants were white European. We found a *TLR7* variant in one of 56 cases (1.7%) and in none of 27 controls. Overall, the association between the presence of *TLR7* rare variants and severe COVID-19 was significant (p=0.037 by Fisher Exact test, ***Table 1***).

We then investigated the presence of *TLR7* rare variants in the entire male cohort of 561 COVID-19 patients (261 cases and 300 controls) regardless of age. We found *TLR7* rare missense variants in three additional patients over 60 years of age, including two cases (who shared the p.Ala1032Thr variant) and one control (C1), bearing the p.Val222Asp variant, predicted to have a low impact on protein function (CADD of 5.36) (***Table 2***).

In order to functionally link the presence of the identified *TLR7* missense variants and the effect on the downstream type I IFN-signaling, we performed a gene expression profile analysis in peripheral blood mononuclear cells (PBMCs) isolated from patients following recovery, after stimulation with the TLR7 agonist imiquimod, as reported by ***van der Made et al., 2020***. To explore all *TLR7* variants identified, we examined PBMCs from the control and all cases except P4 and P6 because them were not available. However, P4 and P5 shared the same variant. This analysis showed a statistically significant decrease of all *TLR7*-related genes for two variants (Ser301Pro and Ala1032Thr) identified in cases P3, P7, and P8 compared with healthy controls (Ctl) demonstrating a complete impairment of TLR7 signaling pathways in response to TLR7 stimulation (***Figure 2***, panel A and ***Table 2***). The variant Val219Ile (P1) showed a hypomorphic effect determining a statistically

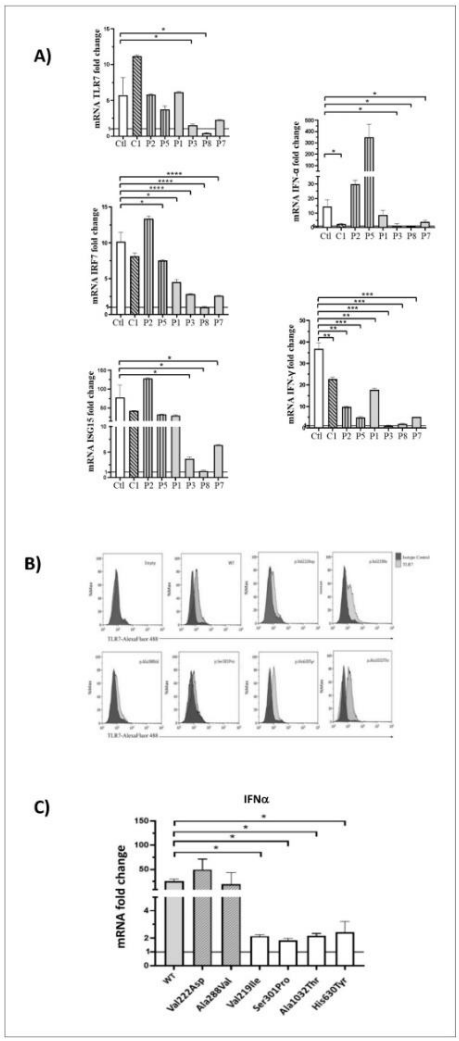


Figure 2. Gene expression profile analysis in peripheral blood mononuclear cells (PBMCs) and in HEK293 cells transfected with the functional variants after stimulation with a TLR7 agonist for 4 hr. (**A**) 5×10^5 PBMCs from COVID-19 patients and six unaffected male and female controls were stimulated for 4 hr with the TLR7 agonist imiquimod at 5 µg/mL or cell culture medium. Quantitative PCR assay was performed and the $2^{-\Delta\Delta Ct}$ calculated using *HPRT1* as housekeeping gene. Fold change in mRNA expression of *TLR7* and type 1 IFN-related genes *ISG15*, *IRF7*, *IFN-α* and *IFN-γ* induced by TLR7

Figure 2 continued on next page

Figure 2 continued

agonist imiquimod was compared with cell culture medium. Ctl indicates healthy controls (white bar); C_1, the asymptomatic mutated control (diagonal lines bar); P2, P5, cases with neutral variants (vertical lines bar); P1, P3, P8, P7 cases with functional variants (gray bar) (as in ***Table 2***). (**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. Quantitative PCR assay was performed and the $2^{-\Delta\Delta Ct}$ calculated using HPRT1 as housekeeping gene. Fold change in mRNA expression induced by imiquimod was compared with cell culture medium. Error bars show standard deviation. p *values* were calculated for the reduction using an unpaired *t* test: *p<0.05; **p<0.01; ***p<0.001; ****p<0.0001.

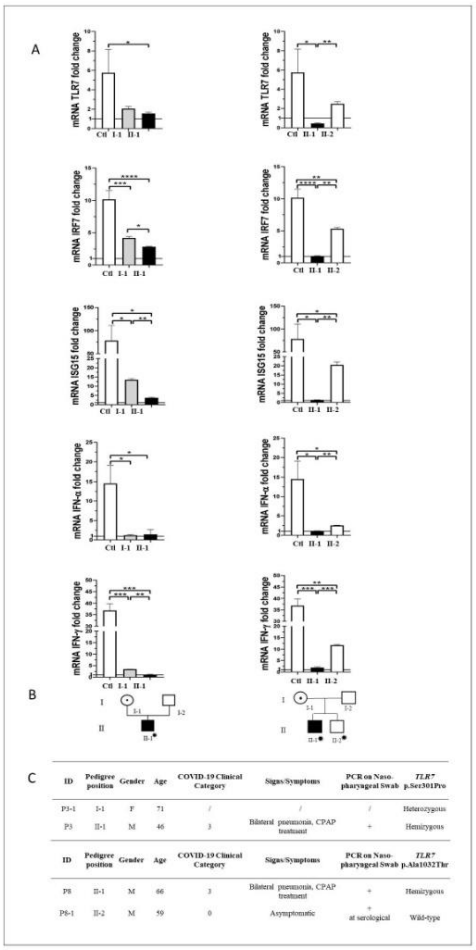
significant decrease in mRNA levels only for IRF7 (directly activated by TLR7) and IFN-γ (***Figure 2***, panel A). Two Ala to Val variants identified in severely affected patients, Ala288Val and Ala448Val, were functionally neutral, that is not predicted to impair the TLR7 signaling pathways. This was confirmed by biochemical and structural analysis on the crystal structure of TLR7 protein (https://www.uniprot.org/uniprot/Q9NYK1). The prediction performed with different computational approaches showed both variants as benign with no effects on structural stabilization. Interestingly, the p.Val222Asp variant (C1) proved to be functionally neutral, in keeping with it being identified in the control and not in cases (***Figure 2***, panel A).

TLR7 expression was evaluated in monocytes and B cells from patients and healthy controls by flow cytometry. Patients and controls expressed the TLR7 protein at the intracellular level. The functional capacity of PBMCs was evaluated after stimulation with the TLR4 agonist lipopolysaccharide (LPS). Of note, LPS-induced production of IL6 by monocytes was similar in patients and controls (data not shown).

In order to validate the functional effect of *TLR7* variants, we have performed transfection experiments in HEK293 cells, cloning a dedicated TLR7 plasmid for each of them. Transfection experiments were performed in HEK293 cells that do not express endogenous TLR7 (***Chehadeh and Alkhabbaz, 2013***) and expression of TLR7 protein was examined by flow cytometry 24 hr after transfection, showing expression of TLR7 protein at the intracellular level in all cases (***Figure 2***, panel B). We then evaluated the expression of IFN-a in imiquimod stimulated and unstimulated cells by qRT-PCR employing the same assay described for PBMCs, confirming the results obtained in PBMCs for the screened variants (***Figure 2***, panel C).

Segregation analysis was available for two cases, P3 and P8 (***Figure 3***). In the two pedigrees, the disease nicely segregated as an X-linked disorder conditioned by environmental factors, that is SARS-CoV-2 (***Figure 3***, panel B). This was also supported by functional analysis on all *TLR7*-related genes (***Figure 3***, panel A). For example, expression profile analysis for *IRF7* gene in male mutated patient P8 confirmed a statistically significant reduction compared to the wild-type brother (***Figure 3***, panel A). Of note, only the infected mutated male had severe COVID-19, whereas the infected not mutated brother (II-2 of P8) was asymptomatic (***Figure 3***, panel C).

Our results showed that the two families reported by ***van der Made et al., 2020***. with loss-of-function variants in males with severe COVID-19 with a mean age of 26 years represent a subset of COVID-19 male patients. 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 (3/135, 2.2%). The same result was obtained for the entire male cohort, irrespective of age, with *TLR7* deleterious variants in 5/261 cases (1.9%). Since not all identified variants were functionally effective, the true percentage could be slightly lower in young males. Overall, males with rare missense variants shown here developed COVID-19 at a mean age of 56.5 years, considerably later than 26 years, in agreement with a predicted smaller impact on the protein than the loss of function deleterious variants reported by ***van der Made et al., 2020***. Similarly, the identified rare missense *TLR7* variants impaired the mRNA expression of *TLR7* as well as the downstream pathway. The observation reported here may lead to consider *TLR7* screening in severely affected male patients in order to start personalized interferon treatment for those with this specific genetic disorder.



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 3. 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**. Gray columns represent individuals harboring 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 ( ).

Acknowledgements

This study is part of the GEN-COVID Multicenter Study, https://sites.google.com/dbm.unisi.it/gen-covid, the Italian multicenter study aimed at identifying the COVID-19 host genetic bases. Specimens were provided by the COVID-19 Biobank of Siena, which is part of the Genetic Biobank of Siena, member of BBMRI-IT, of Telethon Network of Genetic Biobanks (project no. GTB18001), of EuroBioBank, and of RD-Connect. We thank the CINECA consortium for providing computational resources and the Network for Italian Genomes (NIG) http://www.nig.cineca.it for its support. We thank private donors for the support provided to AR (Department of Medical Biotechnologies, University of Siena) for the COVID-19 host genetics research project (D.L n.18 of March 17, 2020). We also thank the COVID-19 Host Genetics Initiative (https://www.covid19hg.org/), MIUR project 'Dipartimenti di Eccellenza 2018–2020' to the Department of Medical Biotechnologies University of Siena, Italy, and 'Bando Ricerca COVID-19 Toscana' project to Azienda Ospedaliero-Universitaria Senese. We also thank Intesa San Paolo for the 2020 charity fund dedicated to the project N B/2020/0119 'Identificazione delle basi genetiche determinanti la variabilità clinica della risposta a COVID-19 nella popolazione italiana'.

Additional information

Group author details

GEN-COVID Multicenter Study

Floriana Valentino: Medical Genetics, University of Siena, Siena, Italy; Med Biotech Hub and Competence Center, Department of Medical Biotechnologies, University of Siena, Siena, Italy; Gabriella Doddato: Medical Genetics, University of Siena, Siena, Italy; Med Biotech Hub and Competence Center, Department of Medical Biotechnologies, University of Siena, Siena, Italy; Annarita Giliberti: Medical Genetics, University of Siena, Siena, Italy; Med Biotech Hub and Competence Center, Department of Medical Biotechnologies, University of Siena, Siena, Italy; Rossella Tita: Genetica Medica, Azienda Ospedaliero-Universitaria Senese, Siena, Italy; Sara Amitrano: Genetica Medica, Azienda Ospedaliero-Universitaria Senese, Siena, Italy; Mirella Bruttini: Medical Genetics, University of Siena, Siena, Italy; Med Biotech Hub and Competence Center, Department of Medical Biotechnologies, University of Siena, Siena, Italy; Genetica Medica, Azienda Ospedaliero-Universitaria Senese, Siena, Italy; Susanna Croci: Medical Genetics, University of Siena, Siena, Italy; Med Biotech Hub and Competence Center, Department of Medical Biotechnologies, University of Siena, Siena, Italy; Ilaria Meloni: Medical Genetics, University of Siena, Siena, Italy; Med Biotech Hub and Competence Center, Department of Medical Biotechnologies, University of Siena, Siena, Italy; Maria Antonietta Mencarelli: Genetica Medica, Azienda Ospedaliero-Universitaria Senese, Siena, Italy; Caterina Lo Rizzo: Genetica Medica, Azienda Ospedaliero-Universitaria Senese, Siena, Italy; Anna Maria Pinto: Genetica Medica, Azienda Ospedaliero-Universitaria Senese, Siena, Italy; Laura Di Sarno: Medical Genetics, University of Siena, Siena, Italy; Med Biotech Hub and Competence Center, Department of Medical Biotechnologies, University of Siena, Siena, Italy; Giada Beligni: Medical Genetics, University of Siena, Siena, Italy; Med Biotech Hub and Competence Center, Department of Medical Biotechnologies, University of Siena, Siena, Italy; Andrea Tommasi: Medical Genetics, University of Siena, Siena, Italy; Med Biotech Hub and Competence Center, Department of Medical Biotechnologies, University of Siena, Siena, Italy; Genetica Medica, Azienda Ospedaliero-Universitaria Senese, Siena, Italy; Nicola Iuso: Medical Genetics, University of Siena, Siena, Italy; Med Biotech Hub and Competence Center, Department of Medical Biotechnologies, University of Siena, Siena, Italy; Francesca Montagnani: Med Biotech Hub and Competence Center, Department of Medical Biotechnologies, University of Siena, Siena, Italy; Dept of Specialized and Internal Medicine, Tropical and Infectious Diseases Unit, Azienda Ospedaliera Universitaria Senese, Siena, Italy; Massimiliano Fabbiani: Dept of Specialized and Internal Medicine, Tropical and Infectious Diseases Unit, Azienda Ospedaliera Universitaria Senese, Siena, Italy; Barbara Rossetti: Dept of Specialized and Internal Medicine, Tropical and Infectious Diseases Unit, Azienda Ospedaliera Universitaria Senese, Siena, Italy; Giacomo Zanelli: Med Biotech Hub and Competence Center, Department of Medical Biotechnologies, University

of Siena, Siena, Italy; Dept of Specialized and Internal Medicine, Tropical and Infectious Diseases Unit, Azienda Ospedaliera Universitaria Senese, Siena, Italy; Elena Bargagli: Unit of Respiratory Diseases and Lung Transplantation, Department of Internal and Specialist Medicine, University of Siena, Siena, Italy; Laura Bergantini: Unit of Respiratory Diseases and Lung Transplantation, Department of Internal and Specialist Medicine, University of Siena, Siena, Italy; Miriana D'Alessandro: Unit of Respiratory Diseases and Lung Transplantation, Department of Internal and Specialist Medicine, University of Siena, Siena, Italy; Paolo Cameli: Unit of Respiratory Diseases and Lung Transplantation, Department of Internal and Specialist Medicine, University of Siena, Siena, Italy; David Bennett: Unit of Respiratory Diseases and Lung Transplantation, Department of Internal and Specialist Medicine, University of Siena, Siena, Italy; Federico Anedda: Dept of Emergency and Urgency, Medicine, Surgery and Neurosciences, Unit of Intensive Care Medicine, Siena University Hospital, Siena, Italy; Simona Marcantonio: Dept of Emergency and Urgency, Medicine, Surgery and Neurosciences, Unit of Intensive Care Medicine, Siena University Hospital, Siena, Italy; Sabino Scolletta: Dept of Emergency and Urgency, Medicine, Surgery and Neurosciences, Unit of Intensive Care Medicine, Siena University Hospital, Siena, Italy; Federico Franchi: Dept of Emergency and Urgency, Medicine, Surgery and Neurosciences, Unit of Intensive Care Medicine, Siena University Hospital, Siena, Italy; Maria Antonietta Mazzei: Department of Medical, Surgical and Neuro Sciences and Radiological Sciences, Unit of Diagnostic Imaging, University of Siena, Siena, Italy; Susanna Guerrini: Department of Medical, Surgical and Neuro Sciences and Radiological Sciences, Unit of Diagnostic Imaging, University of Siena, Siena, Italy; Edoardo Conticini: Rheumatology Unit, Department of Medicine, Surgery and Neurosciences, University of Siena, Policlinico Le Scotte, Siena, Italy; Luca Cantarini: Rheumatology Unit, Department of Medicine, Surgery and Neurosciences, University of Siena, Policlinico Le Scotte, Siena, Italy; Bruno Frediani: Rheumatology Unit, Department of Medicine, Surgery and Neurosciences, University of Siena, Policlinico Le Scotte, Siena, Italy; Danilo Tacconi: Department of Specialized and Internal Medicine, Infectious Diseases Unit, San Donato Hospital Arezzo, San Donato Hospital Arezzo, Arezzo, Italy; Chiara Spertilli: Department of Specialized and Internal Medicine, Infectious Diseases Unit, San Donato Hospital Arezzo, San Donato Hospital Arezzo, Arezzo, Italy; Marco Feri: Dept of Emergency, Anesthesia Unit, San Donato Hospital, Arezzo, Italy; Alice Donati: Dept of Emergency, Anesthesia Unit, San Donato Hospital, Arezzo, Italy; Raffaele Scala: Department of Specialized and Internal Medicine, Pneumology Unit and UTIP, San Donato Hospital, Arezzo, Italy; Luca Guidelli: Department of Specialized and Internal Medicine, Pneumology Unit and UTIP, San Donato Hospital, Arezzo, Italy; Genni Spargi: Department of Emergency, Anesthesia Unit, Misericordia Hospital, Grosseto, Italy; Marta Corridi: Department of Emergency, Anesthesia Unit, Misericordia Hospital, Grosseto, Italy; Cesira Nencioni: Department of Specialized and Internal Medicine, Infectious Diseases Unit, Misericordia Hospital, Grosseto, Italy; Leonardo Croci: Department of Specialized and Internal Medicine, Infectious Diseases Unit, Misericordia Hospital, Grosseto, Italy; Gian Piero Caldarelli: Clinical Chemical Analysis Laboratory, Misericordia Hospital, Grosseto, Italy; Maurizio Spagnesi: Department of Preventive Medicine, Azienda USL Toscana Sud Est, Siena, Italy; Davide Romani: Department of Preventive Medicine, Azienda USL Toscana Sud Est, Siena, Italy; Paolo Piacentini: Department of Preventive Medicine, Azienda USL Toscana Sud Est, Siena, Italy; Elena Desanctis: Department of Preventive Medicine, Azienda USL Toscana Sud Est, Siena, Italy; Silvia Cappelli: Department of Preventive Medicine, Azienda USL Toscana Sud Est, Siena, Italy; Anna Canaccini: Territorial Scientific Technician Department, Azienda USL Toscana Sud Est, Siena, Italy; Agnese Verzuri: Territorial Scientific Technician Department, Azienda USL Toscana Sud Est, Siena, Italy; Valentina Anemoli: Territorial Scientific Technician Department, Azienda USL Toscana Sud Est, Siena, Italy; Agostino Ognibene: Clinical Chemical Analysis Laboratory, San Donato Hospital, Arezzo, Italy; Antonella D'Arminio Monforte: Department of Health Sciences, Clinic of Infectious Diseases, ASST Santi Paolo e Carlo, University of Milan, Milano, Italy; Federica Gaia Miraglia: Department of Health Sciences, Clinic of Infectious Diseases, ASST Santi Paolo e Carlo, University of Milan, Milano, Italy; Massimo Girardis: Department of Anesthesia and Intensive Care, University of Modena and Reggio Emilia, Modena, Italy; Sophie Venturelli: Department of Anesthesia and Intensive Care, University of Modena and Reggio Emilia, Modena, Italy; Stefano Busani: Department of Anesthesia and Intensive Care, University of Modena and Reggio Emilia, Modena, Italy; Andrea

Cossarizza: Department of Medical and Surgical Sciences for Children and Adults, University of Modena and Reggio Emilia, Modena, Italy; Andrea Antinori: HIV/AIDS Department, National Institute for Infectious Diseases, IRCCS, Lazzaro Spallanzani, Rome, Italy; Alessandra Vergori: HIV/AIDS Department, National Institute for Infectious Diseases, IRCCS, Lazzaro Spallanzani, Rome, Italy; Arianna Emiliozzi: HIV/AIDS Department, National Institute for Infectious Diseases, IRCCS, Lazzaro Spallanzani, Rome, Italy; Arianna Gabrieli: Department of Biomedical and Clinical Sciences Luigi Sacco, University of Milan, Milan, Italy; Agostino Riva: III Infectious Diseases Unit, ASST-FBF-Sacco, Milan, Italy; Department of Biomedical and Clinical Sciences Luigi Sacco, University of Milan, Milan, Italy; Pier Giorgio Scotton: Department of Infectious Diseases, Treviso Hospital, Local Health Unit 2 Marca Trevigiana, Treviso, Italy; Francesca Andretta: Department of Infectious Diseases, Treviso Hospital, Local Health Unit 2 Marca Trevigiana, Treviso, Italy; Sandro Panese: Clinical Infectious Diseases, Mestre Hospital, Venezia, Italy; Renzo Scaggiante: Infectious Diseases Clinic, ULSS1, Belluno, Italy; Francesca Gatti: Infectious Diseases Clinic, ULSS1, Belluno, Italy; Saverio Giuseppe Parisi: Department of Molecular Medicine, University of Padova, Padua, Italy; Stefano Baratti: Department of Molecular Medicine, University of Padova, Padua, Italy; Melania Degli Antoni: Department of Infectious and Tropical Diseases, University of Brescia and ASST Spedali Civili Hospital, Brescia, Italy; Matteo Della Monica: Medical Genetics and Laboratory of Medical Genetics Unit, A.O.R.N. "Antonio Cardarelli", Naples, Italy; Carmelo Piscopo: Medical Genetics and Laboratory of Medical Genetics Unit, A.O.R.N. "Antonio Cardarelli", Naples, Italy; Mario Capasso: Department of Molecular Medicine and Medical Biotechnology, University of Naples Federico II, Naples, Italy; CEINGE Biotecnologie Avanzate, Naples, Italy; IRCCS SDN, Naples, Italy; Roberta Russo: Department of Molecular Medicine and Medical Biotechnology, University of Naples Federico II, Naples, Italy; CEINGE Biotecnologie Avanzate, Naples, Italy; Immacolata Andolfo: Department of Molecular Medicine and Medical Biotechnology, University of Naples Federico II, Naples, Italy; CEINGE Biotecnologie Avanzate, Naples, Italy; Achille Iolascon: Department of Molecular Medicine and Medical Biotechnology, University of Naples Federico II, Naples, Italy; CEINGE Biotecnologie Avanzate, Naples, Italy; Giuseppe Fiorentino: Unit of Respiratory Physiopathology, AORN dei Colli, Monaldi Hospital, Naples, Italy; Massimo Carella: Division of Medical Genetics, Fondazione IRCCS Casa Sollievo della Sofferenza Hospital, San Giovanni Rotondo, San Giovanni Rotondo, Italy; Marco Castori: Division of Medical Genetics, Fondazione IRCCS Casa Sollievo della Sofferenza Hospital, San Giovanni Rotondo, San Giovanni Rotondo, Italy; Giuseppe Merla: Department of Molecular Medicine and Medical Biotechnology, University of Naples Federico II, Naples, Italy; Laboratory of Regulatory and Functional Genomics, Fondazione IRCCS Casa Sollievo della Sofferenza, San Giovanni Rotondo, Italy; Gabriella Maria Squeo: Laboratory of Regulatory and Functional Genomics, Fondazione IRCCS Casa Sollievo della Sofferenza, San Giovanni Rotondo, Italy; Filippo Aucella: Department of Medical Sciences, Fondazione IRCCS Casa Sollievo della Sofferenza Hospital, San Giovanni Rotondo, San Giovanni Rotondo, Italy; Pamela Raggi: Clinical Trial Office, Fondazione IRCCS Casa Sollievo della Sofferenza Hospital, San Giovanni Rotondo, San Giovanni Rotondo, Italy; Carmen Marciano: Clinical Trial Office, Fondazione IRCCS Casa Sollievo della Sofferenza Hospital, San Giovanni Rotondo, San Giovanni Rotondo, Italy; Rita Perna: Clinical Trial Office, Fondazione IRCCS Casa Sollievo della Sofferenza Hospital, San Giovanni Rotondo, San Giovanni Rotondo, Italy; Matteo Bassetti: Department of Health Sciences, University of Genova, Genova, Italy; Infectious Diseases Clinic, Policlinico San Martino Hospital, IRCCS for Cancer Research Genova, Genova, Italy; Antonio Di Biagio: Infectious Diseases Clinic, Policlinico San Martino Hospital, IRCCS for Cancer Research Genova, Genova, Italy; Maurizio Sanguinetti: Microbiology, Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Catholic University of Medicine, Rome, Italy; Department of Laboratory Sciences and Infectious Diseases, Fondazione Policlinico Universitario A. Gemelli IRCCS, Rome, Italy; Luca Masucci: Microbiology, Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Catholic University of Medicine, Rome, Italy; Department of Laboratory Sciences and Infectious Diseases, Fondazione Policlinico Universitario A. Gemelli IRCCS, Rome, Italy; Serafina Valente: Department of Cardiovascular Diseases, University of Siena, Siena, Italy; Marco Mandalà: Otolaryngology Unit, University of Siena, Siena, Italy; Alessia Giorli: Otolaryngology Unit, University of Siena, Siena, Italy; Lorenzo Salerni: Otolaryngology Unit, University of Siena, Siena, Italy; Patrizia Zucchi: Department of Internal Medicine, ASST Valtellina e Alto Lario, Sondrio, Italy; Pierpaolo

Parravicini: Department of Internal Medicine, ASST Valtellina e Alto Lario, Sondrio, Italy; Elisabetta Menatti: Study Coordinator Oncologia Medica e Ufficio Flussi Sondrio, Sondrio, Italy; Tullio Trotta: First Aid Department, Luigi Curto Hospital, Polla, Salerno, Italy; Ferdinando Giannattasio: First Aid Department, Luigi Curto Hospital, Polla, Salerno, Italy; Gabriella Coiro: First Aid Department, Luigi Curto Hospital, Polla, Salerno, Italy; Fabio Lena: Local Health Unit-Pharmaceutical Department of Grosseto, Toscana Sud Est Local Health Unit, Grosseto, Italy; Domenico A Coviello: U.O.C. Laboratorio di Genetica Umana, IRCCS Istituto G. Gaslini, Genova, Italy; Cristina Mussini: Infectious Diseases Clinics, University of Modena and Reggio Emilia, Modena, Italy; Giancarlo Bosio: Department of Respiratory Diseases, Azienda Ospedaliera di Cremona, Cremona, Italy; Enrico Martinelli: Department of Respiratory Diseases, Azienda Ospedaliera di Cremona, Cremona, Italy; Sandro Mancarella: U.O.C. Medicina, ASST Nord Milano, Ospedale Bassini, Cinisello Balsamo, Italy; Luisa Tavecchia: U.O.C. Medicina, ASST Nord Milano, Ospedale Bassini, Cinisello Balsamo, Italy; Marco Gori: Université Côte d'Azur, Inria, France; Lia Crotti: Istituto Auxologico Italiano, IRCCS, Department of Cardiovascular, Neural and Metabolic Sciences, San Luca Hospital, Milan, Italy; Department of Medicine and Surgery, University of Milano-Bicocca, Milan, Italy; Istituto Auxologico Italiano, IRCCS, Center for Cardiac Arrhythmias of Genetic Origin, Milan, Italy; Istituto Auxologico Italiano, IRCCS, Laboratory of Cardiovascular Genetics, Milan, Italy; Member of the European Reference Network for Rare, Low Prevalence and Complex Diseases of the Heart-ERN GUARD-Heart; Gianfranco Parati: Istituto Auxologico Italiano, IRCCS, Department of Cardiovascular, Neural and Metabolic Sciences, San Luca Hospital, Milan, Italy; Department of Medicine and Surgery, University of Milano-Bicocca, Milan, Italy; Chiara Gabbi: Independent Medical Scientist, Milan, Italy; Isabella Zanella: Department of Molecular and Translational Medicine, University of Brescia, Brescia, Italy; Clinical Chemistry Laboratory, Cytogenetics and Molecular Genetics Section, Diagnostic Department, ASST Spedali Civili di Brescia, Brescia, Italy; Marco Rizzi: Unit of Infectious Diseases, ASST Papa Giovanni XXIII Hospital, Bergamo, Italy; Franco Maggiolo: Unit of Infectious Diseases, ASST Papa Giovanni XXIII Hospital, Bergamo, Italy; Diego Ripamonti: Unit of Infectious Diseases, ASST Papa Giovanni XXIII Hospital, Bergamo, Italy; Tiziana Bachetti: Direzione Scientifica, Istituti Clinici Scientifici Maugeri IRCCS, Pavia, Italy; Maria Teresa La Rovere: Istituti Clinici Scientifici Maugeri IRCCS, Department of Cardiology, Institute of Montescano, Pavia, Italy; Simona Sarzi-Braga: Istituti Clinici Scientifici Maugeri, IRCCS, Department of Cardiac Rehabilitation, Institute of Tradate (VA), Tradate, Italy; Maurizio Bussotti: Istituti Clinici Scientifici Maugeri, IRCCS, Department of Cardiac Rehabilitation, Institute of Milan, Milan, Italy; Mario Chiariello: Istituto per lo Studio, la Prevenzione e la Rete Oncologica (ISPRO)-Core Research Laboratory and Consiglio Nazionale delle Ricerche-Istituto di Fisiologia Clinica, Siena, Italy; Mary Ann Belli: ASST Nord Milano, Ospedale Bassini, Cinisello Balsamo, Italy; Simona Dei: Health Management, Azienda USL Toscana Sudest, Tuscany, Italy

Funding

Funder	Grant reference number	Author
Private Donors for Host Genetics Research Project	D.L. n 18 of March 17	Alessandra Renieri
Intesa San Paolo for 2020 charity fund	N.B.2020/0119	Alessandra Renieri
Ministero dell'Istruzione, dell'Università e della Ricerca	Dipartimenti di Eccellenza 2018-2020	Alessandra Renieri
Regione Toscana	Bando Ricerca COVID-19 Toscana	Alessandra Renieri

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Chiara Fallerini, Formal analysis, Writing - original draft; Sergio Daga, Margherita Baldassarri, Francesca Fava, Katia Capitani, Formal analysis, Methodology, Writing - original draft; Stefania Mantovani, Daniela Francisci, Francesco Paciosi, Elisabetta Schiaroli, Maria Palmieri, Serena Ludovisi,

Francesco Castelli, Eugenia Quiros-Roldan, Massimo Vaghi, Stefano Rusconi, Matteo Siano, Maria Bandini, Mario U Mondelli, Methodology, Writing - original draft; Elisa Benetti, Software, Formal analysis, Methodology, Writing - original draft; Nicola Picchiotti, Software, Methodology, Writing - original draft; Ottavia Spiga, GEN-COVID Multicenter Study, Data curation, Formal analysis, Methodology, Writing - original draft; Simone Furini, Data curation, Software, Formal analysis, Supervision, Validation, Methodology, Writing - original draft; Francesca Mari, Data curation, Methodology, Writing - original draft, Project administration; GEN-COVID Multicenter Study, Conceptualization, Methodology, Writing - original draft; Alessandra Renieri, Conceptualization, Data curation, Supervision, Writing - original draft, Project administration; Elisa Frullanti, Conceptualization, Data curation, Formal analysis, Supervision, Methodology, Writing - original draft, Project administration

Author ORCIDs
Chiara Fallerini https://orcid.org/0000-0002-7386-3224
Sergio Daga https://orcid.org/0000-0002-6419-9456
Stefania Mantovani https://orcid.org/0000-0002-5885-2842
Elisa Benetti https://orcid.org/0000-0002-0819-604X
Nicola Picchiotti https://orcid.org/0000-0003-3454-7250
Margherita Baldassarri https://orcid.org/0000-0002-0391-1980
Francesca Fava https://orcid.org/0000-0002-4363-2353
Maria Palmieri https://orcid.org/0000-0003-3014-1552
Simone Furini https://orcid.org/0000-0002-1099-8279
Alessandra Renieri https://orcid.org/0000-0002-0846-9220
Elisa Frullanti https://orcid.org/0000-0001-5634-031X

Ethics
Clinical trial registration NCT04549831.
Human subjects: The GEN-COVID study was consistent with Institutional guidelines and approved by the University Hospital (Azienda Ospedaliero-Universitaria Senese) Ethical Review Board, Siena, Italy (Prot n. 16929, dated March 16, 2020).

Decision letter and Author response
Decision letter https://doi.org/10.7554/eLife.67569.sa1
Author response https://doi.org/10.7554/eLife.67569.sa2

Additional files

Supplementary files
- Reporting standard 1. STROBE checklist.
- Transparent reporting form

Data availability
Sequencing data have been deposited in CINECA through http://www.nig.cineca.it/, specifically, http://nigdb.cineca.it., in the COVID-19 section through http://nigdb.cineca.it./registration/login.php. There are no restrictions on data access. Only registration is needed.

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nature Methods* **7**:248–249. DOI: https://doi.org/10.1038/nmeth0410-248, PMID: 20354512

Bastard P, Rosen LB, Zhang Q, Michailidis E, Hoffmann HH, Zhang Y, Dorgham K, Philippot Q, Rosain J, Béziat V, Manry J, Shaw E, Haljasmägi L, Peterson P, Lorenzo L, Bizien L, Trouillet-Assant S, Dobbs K, de Jesus AA, Belot A, et al. 2020. Autoantibodies against type I IFNs in patients with life-threatening COVID-19. *Science* **370**:eabd4585. DOI: https://doi.org/10.1126/science.abd4585, PMID: 32972996

Chehadeh W, Alkhabbaz M. 2013. Differential TLR7-mediated expression of proinflammatory and antiviral cytokines in response to laboratory and clinical Enterovirus strains. *Virus Research* **174**:88–94. DOI: https://doi.org/10.1016/j.virusres.2013.03.006, PMID: 23523654

Daga S, Fallerini C, Baldassarri M, Fava F, Valentino F, Doddato G, Benetti E, Furini S, Giliberti A, Tita R, Amitrano S, Bruttini M, Meloni I, Pinto AM, Raimondi F, Stella A, Biscarini F, Picchiotti N, Gori M, Pinoli P, et al. 2021. Employing a systematic approach to biobanking and analyzing clinical and genetic data for advancing COVID-19 research. *European Journal of Human Genetics* **1**:1–15. DOI: https://doi.org/10.1038/s41431-020-00793-7

Garg S, Kim L, Whitaker M, O'Halloran A, Cummings C, Holstein R, Prill M, Chai SJ, Kirley PD, Alden NB, Kawasaki B, Yousey-Hindes K, Niccolai L, Anderson EJ, Openo KP, Weigel A, Monroe ML, Ryan P, Henderson J, Kim S, et al. 2020. Hospitalization rates and characteristics of patients hospitalized with Laboratory-Confirmed coronavirus disease 2019 - COVID-NET, 14 states, march 1-30, 2020. *MMWR. Morbidity and Mortality Weekly Report* **69**:458–464. DOI: https://doi.org/10.15585/mmwr.mm6915e3, PMID: 32298251

Goodman KE, Magder LS, Baghdadi JD, Pineles L, Levine AR, Perencevich EN, Harris AD. 2020. Impact of sex and metabolic comorbidities on COVID-19 mortality risk across age groups: 66,646 inpatients across 613 U.S. hospitals. *Clinical Infectious Diseases* **18**:ciaa1787. DOI: https://doi.org/10.1093/cid/ciaa1787

Hägg S, Jylhävä J, Wang Y, Xu H, Metzner C, Annetorp M, Garcia-Ptacek S, Khedri M, Boström AM, Kadir A, Johansson A, Kivipelto M, Eriksdotter M, Cederholm T, Religa D. 2020. Age, frailty, and comorbidity as prognostic factors for Short-Term outcomes in patients with coronavirus disease 2019 in geriatric care. *Journal of the American Medical Directors Association* **21**:1555–1559. DOI: https://doi.org/10.1016/j.jamda.2020.08.014, PMID: 32978065

Iavarone C, Ramsauer K, Kubarenko AV, Debasitis JC, Leykin I, Weber AN, Siggs OM, Beutler B, Zhang P, Otten G, D'Oro U, Valiante NM, Mbow ML, Visintin A. 2011. A point mutation in the amino terminus of TLR7 abolishes signaling without affecting ligand binding. *The Journal of Immunology* **186**:4213–4222. DOI: https://doi.org/10.4049/jimmunol.1003585, PMID: 21383246

Mantovani S, Oliviero B, Lombardi A, Varchetta S, Mele D, Sangiovanni A, Rossi G, Donadon M, Torzilli G, Soldani C, Porta C, Pedrazzoli P, Chiellino S, Santambrogio R, Opocher E, Maestri M, Bernuzzi S, Rossello A, Clément S, De Vito C, et al. 2019. Deficient natural killer cell NKp30-Mediated function and altered NCR3 splice variants in hepatocellular carcinoma. *Hepatology* **69**:1165–1179. DOI: https://doi.org/10.1002/hep.30235, PMID: 30153337

Mele D, Mantovani S, Oliviero B, Grossi G, Ludovisi S, Mondelli MU, Varchetta S. 2017. Hepatitis C virus inhibits CD4 T cell function via binding to Toll-like receptor 7. *Antiviral Research* **137**:108–111. DOI: https://doi.org/10.1016/j.antiviral.2016.11.013

Ng PC, Henikoff S. 2003. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research* **31**:3812–3814. DOI: https://doi.org/10.1093/nar/gkg509, PMID: 12824425

Poulas K, Farsalinos K, Zanidis C. 2020. Activation of TLR7 and innate immunity as an efficient method against COVID-19 pandemic: imiquimod as a potential therapy. *Frontiers in Immunology* **11**:1373. DOI: https://doi.org/10.3389/fimmu.2020.01373, PMID: 32612613

Rodrigues CH, Pires DE, Ascher DB. 2018. DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic Acids Research* **46**:W350–W355. DOI: https://doi.org/10.1093/nar/gky300, PMID: 29718330

Stokes EK, Zambrano LD, Anderson KN, Marder EP, Raz KM, El Burai Felix S, Tie Y, Fullerton KE. 2020. Coronavirus disease 2019 case surveillance - United states, January 22-May 30, 2020. *MMWR. Morbidity and Mortality Weekly Report* **69**:759–765. DOI: https://doi.org/10.15585/mmwr.mm6924e2, PMID: 32555134

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B* **58**:267–288. DOI: https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

van der Made CI, Simons A, Schuurs-Hoeijmakers J, van den Heuvel G, Mantere T, Kersten S, van Deuren RC, Steehouwer M, van Reijmersdal SV, Jaeger M, Hofste T, Astuti G, Corominas Galbany J, van der Schoot V, van der Hoeven H, Hagmolen of ten Have W, Klijn E, van den Meer C, Fiddelaers J, de Mast Q, et al. 2020. Presence of genetic variants among young men with severe COVID-19. *JAMA* **324**:663–1111. DOI: https://doi.org/10.1001/jama.2020.13719

Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, de Beer TAP, Rempfer C, Bordoli L, Lepore R, Schwede T. 2018. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research* **46**:W296–W303. DOI: https://doi.org/10.1093/nar/gky427, PMID: 29788355

Wu Z, McGoogan JM. 2020. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) Outbreak in China: summary of a report of 72314 Cases From the Chinese Center for Disease Control and Prevention. *Jama* **323**:1239–1242. DOI: https://doi.org/10.1001/jama.2020.2648, PMID: 32091533

Zhang Q, Bastard P, Liu Z, Le Pen J, Moncada-Velez M, Chen J, Ogishi M, Sabli IKD, Hodeib S, Korol C, Rosain J, Bilguvar K, Ye J, Bolze A, Bigio B, Yang R, Arias AA, Zhou Q, Zhang Y, Onodi F, et al. 2020. Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. *Science* **370**:eabd4570. DOI: https://doi.org/10.1126/science.abd4570, PMID: 32972995

7. *SELP* Asp603Asn and severe thrombosis in COVID-19 males

While thromboembolism is a frequent cause of severity and mortality in COVID-19, the etiology of this phenomenon is not well understood. Zang *et* colleagues showed that the SARS-CoV-2 virus directly activates platelets and enhances their prothrombotic function and inflammatory response via binding of Spike to ACE2 [61]. However, why the excess of thromboembolic events happens in some individuals and not in others is still unexplored.

In the present chapter, by applying the LASSO logistic regression model on the Boolean representation of common variants (Recessive model) of autosomal genes (C_AR Boolean described in chapter 2, section 2.5.1) we identified *SELP* as the genetic factor predisposing males to thromboembolism and severe COVID-19 [62]. We also showed that the predisposition increases if the protective effect of testosterone is lost either by age or because of additional genetic factors such as polyQ≥23 in the androgen receptor (AR) gene (presented in Chapter 5).

Fallerini *et al. J Hematol Oncol* (2021) 14:123
https://doi.org/10.1186/s13045-021-01136-9

Journal of Hematology & Oncology

LETTER TO THE EDITOR Open Access

SELP Asp603Asn and severe thrombosis in COVID-19 males

Chiara Fallerini[1,2], Sergio Daga[1,2], Elisa Benetti[2], Nicola Picchiotti[3,4], Kristina Zguro[2], Francesca Catapano[1,2], Virginia Baroni[1,2], Simone Lanini[5], Alessandro Bucalossi[6], Giuseppe Marotta[6], Francesca Colombo[7], Margherita Baldassarri[1,2], Francesca Fava[1,2,8], Giada Beligni[1,2], Laura Di Sarno[1,2], Diana Alaverdian[1,2], Maria Palmieri[1,2], Susanna Croci[1,2], Andrea M. Isidori[9], Simone Furini[2], Elisa Frullanti[1,2] on behalf of GEN-COVID Multicenter Study, Alessandra Renieri[1,2,8*] and Francesca Mari[1,2,8]

Abstract

Thromboembolism is a frequent cause of severity and mortality in COVID-19. However, the etiology of this phenomenon is not well understood. A cohort of 1186 subjects, from the GEN-COVID consortium, infected by SARS-CoV-2 with different severity was stratified by sex and adjusted by age. Then, common coding variants from whole exome sequencing were mined by LASSO logistic regression. The homozygosity of the cell adhesion molecule P-selectin gene (*SELP*) rs6127 (c.1807G > A; p.Asp603Asn) which has been already associated with thrombotic risk is found to be associated with severity in the male subcohort of 513 subjects (odds ratio = 2.27, 95% Confidence Interval 1.54–3.36). As the *SELP* gene is downregulated by testosterone, the odd ratio is increased in males older than 50 (OR 2.42, 95% CI 1.53–3.82). Asn/Asn homozygotes have increased D-dimers values especially when associated with poly Q ≥ 23 in the androgen receptor (OR 3.26, 95% CI 1.41–7.52). These results provide a rationale for the repurposing of antibodies against P-selectin as adjuvant therapy in rs6127 male homozygotes especially if older than 50 or with an impaired androgen receptor.

Keywords: COVID-19, Thromboembolism, Thrombus, Venous thromboembolism, P-selectin, Anti-selectin P monoclonal antibodies

To the Editor

It is now widely recognized that COVID-19 is a systemic disease, characterized by dysregulation of the immune system and by a hypercoagulable state [1]. The bases of this prothrombotic susceptibility remain until now elusive, even if it is evident that host genetic factors largely contribute to COVID-19 phenotypic variability. Rare variants of genes involved in adaptive immunity have been identified in Mendelian forms of COVID-19, where the presence of one rare mutation leads to a severe COVID-19 phenotype segregating in the family following a classic Mendelian inheritance pattern [2]. Among common genetic factors, the protective role of the 0 blood group has been identified, at least in part possibly due to von Willebrand factor (vWF) destabilization protecting from thrombosis [3]. We have also shown that longer polyQ repeats (≥ 23) in the androgen receptor (AR) predispose to severe COVID-19 outcome due to reduced testosterone anti-inflammatory and anti-thrombotic effect [4].

The P-selectin (*SELP*) gene encodes a cell adhesion molecule mediating the interaction of activated platelets on endothelium with leukocytes and playing a key role in thrombosis [5, 6]. Furthermore, significantly increased P-selectin and other prothrombotic biomarkers

*Correspondence: alessandra.renieri@unisi.it
[1] Medical Genetics Unit, University of Siena, Policlinico Le Scotte, Viale Bracci, 2, 53100 Siena, Italy
Full list of author information is available at the end of the article

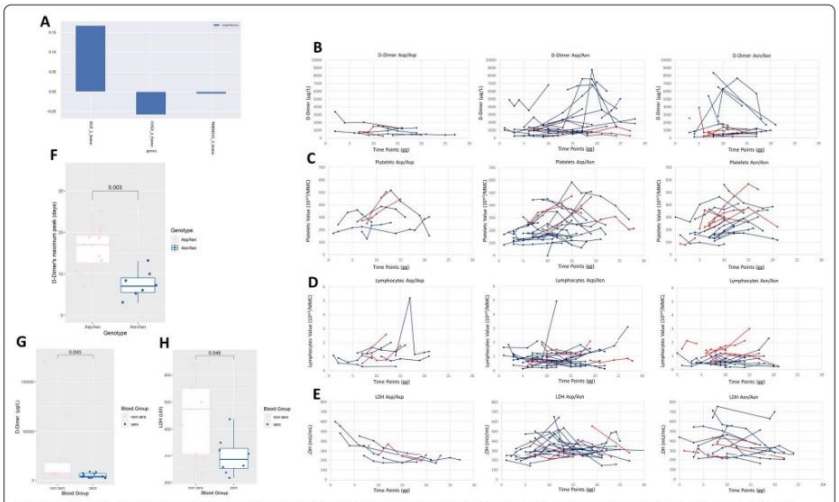


Fig. 1 Homozygous genotype Asn/Asn at the polymorphic locus Asp603Asn (rs6127) is related to severity and to D-dimer pick. **a** Selection of *SELP* gene as relevant for severity. LASSO logistic regression on Boolean representation of homozygous common bi-allelic polymorphism of autosomal genes in males is presented (see paper Picchiotti et al. 2021 for complete representations)[20]. 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 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 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 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 10^3/mmc (**c**), lymphocytes 10^3/mmc (**d**), LDH UI/L (n.v. 135- 225 UI/L) (**e**). As expected, the Asn/Asn homozygous genotype was over-represented (36.53%). Values are reported on the Y-axis. In each graph, the time point "0" (X-axis) represents the day of onset of COVID-19 symptoms. Each line represents each severe hospitalized patient (see methods). 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 years are in red. From left to right patients having Asp/Asp homozygous; Asp/Asn heterozygous; and Asn/Asn homozygous genotype. Older patients only (blue) and Asp/Asn-Asn/Asn genotype only show the D-dimer pick. Accordingly, older patients of these two genotypes have more platelet consumption and higher LDH values. A total of 51 patients have been included in **c**. Among these, 23 patients have a platelet count value below 150×10^3/mmc: 9 with the Asn/Asn genotype, 13 with Asn/Asp and 1 with Asp/Asp. A total of 48 patients have been included in panel D. Among these, 27 patients have lymphocyte count below 0.9 10^3/mmc: 4 Asn/Asn, 19 Asn/Asp and 4 Asp/Asp. A total of 50 patients have been included in panel E. Among these, 44 have LDH values above 225 UI/L: 16 Asn/Asn, 23 Asn/Asp and 5 Asp/Asp. **f** The D-dimer pick is earlier in the Asn/Asn (median = 7.5 days) than the Asp/Asn genotype ($p = 3 \times 10^{-2}$ 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. A total of 47 patients have been included in panel B. Among these, 20 patients show D-Dimer values above 2000 μg/L: 7 Asn/Asn, 12 Asn/Asp and 1 Asp/Asp. **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

concentration in plasma samples of severe COVID-19 patients compared to healthy controls has been recently reported [7, 8].

Among *SELP* variants, the Asp603Asn functional polymorphism (rs6127; c.1807G > A-previously reported as Asp562Asn or Asp541Asn) has been associated with thrombotic risk in various conditions [9, 10]. The polymorphism, together with other coding polymorphisms, 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 [10].

Within the Italian GEN-COVID cohort, we applied an ordered logistic regression to the clinical WHO gradings,

Table 1 Chi-square test in male cohort calculated for all ages (a); for age ≥ 50 years (b); and combination of *AR* poly-Q ≥ 23 and D-dimer value (c)

a	Severe (%)	Mild (%)	Marginal row totals
Chi-square test in male cohort (all ages)			
Asn/Asn genotype	90 (38.14)	59 (21.30)	149
Asp/Asp and Asp/Asn genotype	146 (61.86)	218 (78.70)	364
Marginal column totals	236 (100)	277 (100)	513 (grand total)
b	**Severe (%)**	**Mild (%)**	**Marginal row totals**
Chi-square test in males ≥ 50 years			
Asn/Asn genotype	73 (39.25)	40(21.05)	113
Asp/Asp and Asp/Asn genotype	113 (60.75)	150 (78.95)	263
Marginal column totals	186 (100)	190 (100)	376 (grand total)
c	**D-dimer > 5000**	**D-dimer < 5000**	**Marginal row totals**
Chi-square test of combination of AR poly Q ≥ 23 and D-dimer value			
Asn/Asn and AR polyQ ≥ 23	10	19	29
Asp/Asp and Asp/Asn and AR poliQ < 23	40	248	288
Marginal Column totals	50	267	317 (grand total)

p value (severe vs mild) = 2.8×10^{-5} (OR 2.27, 95% CI 1.54–3.36)

p value (severe vs mild) = 1.19×10^{-4} (OR 2.42, 95% CI 1.53–3.82)

p value (D-dimer > 5000 vs D-dimer < 5000) = 3.73×10^{-3} (OR 3.26, 95% CI 1.41–7.52)

stratified by sex and adjusted by age in order to define severe and mild patients (see Additional file 1: Supplementary file). We then tested by LASSO logistic regression different combinations of coding polymorphisms in homozygous state and found that the *SELP* rs6127 polymorphism correlates with severity only in the subcohort of males (Fig. 1a; Table 1a; Supplementary file; data on females not shown). The genotypic frequencies of the polymorphism in severe and mild patients were confirmed to be in Hardy–Weinberg equilibrium; the minor allele frequency in our cohort was similar to that reported in the European (non-Finnish) population in the gnomAD database (56.2% vs 55.8%) (https://gnomad.broadinstitute.org/).

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 a double copy of Asparagine 603 [10]. These results are in line with the demonstration that SARS-CoV-2 induces thrombosis by binding to ACE2 on platelets and subsequent integrin αIIbβ3 activation and P-selectin expression [11], and that P-selectin soluble isoform is increased in thrombosis [6] and severe COVID-19 [7, 8].

Since *SELP* transcription is inhibited by androgens [12], 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 (Table 1).

In a subset of 52 severely affected hospitalised males, four main laboratory parameters related to a proinflammatory state (lymphocyte count, D-dimer and LDH) and a higher risk for thrombosis (D-dimer, platelet count and LDH) were longitudinally followed (Fig. 1b–e). We observed that the maximum pick (over 10 times of the normal upper value) was exclusive of Asp/Asn and Asn/Asn genotypes and older patients (Fig. 1b–e). 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 $= 3 \times 10^{-2}$, Fig. 1f). As 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 (Fig. 1g, h), in agreement with previous reports [3].

Given the stronger association of the *SELP* polymorphism in older males, the AR poly-Q status would impact on the *SELP* genotype [4]: the combination of poly-Q ≥ 23 with homozygous *SELP* polymorphism versus D-dimer value reached an OR of 3.26 (Table 1c). This result indicates that the two polymorphisms enhance each other, being two pieces of the same puzzle contributing to thrombosis in COVID-19 males.

Anti-*P-Selectin* monoclonal antibodies have been developed for human use: the phase-3 Inclacumab and the FDA&EMA approved Crizanlizumab, the latter as a prevention of vaso-occlusive crises in patients with sickle cell disease [13]. A general clinical trial to test the efficacy and safety of Crizanlizumab in not selected hospitalized COVID-19

patients is ongoing (https://clinicaltrials.gov/ct2/show/study/NCT04435184). Clinical trials in COVID-19 hospitalised males with *SELP* rs6127 should now be encouraged.

Abbreviations
AR: Androgen receptor; SELP: P-selectin gene; vWF: Von Willebrand factor.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13045-021-01136-9.

Additional file 1. Material and Methods plus study group appendix.

Acknowledgements
This study is part of the GEN-COVID Multicenter Study, https://sites.google.com/dbm.unisi.it/gen-covid, the Italian multicenter study aimed at identifying the COVID-19 host genetic bases. Specimens were provided by the COVID-19 Biobank of Siena, which is part of the Genetic Biobank of Siena, member of BBMRI-IT, of Telethon Network of Genetic Biobanks (project no. GTB18001), of EuroBioBank, and of RDConnect. We thank the CINECA consortium for providing computational resources and the Network for Italian Genomes (NIG) http://www.nig.cineca.it for its support. We thank private donors for the support provided to A.R. (Department of Medical Biotechnologies, University of Siena) for the COVID-19 host genetics research project (D.L n.18 of March 17, 2020).

Authors' contributions
AR, FM and EF designed the study; CF, SD, EB, NP, KZ, FC, VB, GB, LDS, DA, SL, SC, MP, AB, GM, AMI, EF, SF analyzed the data; EB, KZ, NP, SF performed statistical analysis; MB, FF and GEN-COVID Multicenter Study provided clinical data; AR and FM supervised the study. All authors read and approved the final manuscript.

Funding
We thank the COVID-19 Host Genetics Initiative (https://www.covid19hg.org/), MIUR project "Dipartimenti di Eccellenza 2018–2020" to the Department of Medical Biotechnologies University of Siena, Italy and "Bando Ricerca COVID-19 Toscana" project to Azienda Ospedaliero Universitaria Senese. We also thank Intesa SanPaolo for the 2020 charity fund dedicated to the project "N. B/2020/0119 Identificazione delle basi genetiche determinanti la variabilità clinica della risposta a COVID-19 nella popolazione italiana"; the Italian Ministry of University and Research for funding within the "Bando FISR 2020" in COVID-19 and the Istituto Buddista Italiano Soka Gakkai for funding the project "PAT-COVID: Host genetics and pathogenetic mechanisms of COVID-19" (ID n. 2020-2016_RIC_3).

Availability of data materials
The data are available for sharing through the COVID-19 dedicated section (http://nigdb.cineca.it), within the Network for Italian Genome (http://www.nig.cineca.it). The data and samples referenced here are housed in the GEN-COVID Patient Registry and the GEN-COVID Biobank and are available for consultation. You may contact the corresponding author, Prof. Alessandra Renieri (e-mail: alessandra.renieri@unisi.it).

Declarations

Ethics approval and consent to participate
The study (GEN-COVID) was consistent with Institutional guidelines and approved by the University Hospital (Azienda Ospedaliero-Universitaria Senese) Ethical Review Board, Siena, Italy (Prot n. 16917, dated March 16, 2020). The patients were informed of this research and agreed to it through the informed consent process.

Consent for publication
Not applicable.

Competing interests
All the authors declare no competing financial interests.

Author details
[1]Medical Genetics Unit, University of Siena, Policlinico Le Scotte, Viale Bracci, 2, 53100 Siena, Italy. [2]Department of Medical Biotechnologies, Med Biotech Hub and Competence Center, University of Siena, Siena, Italy. [3]Department of Mathematics, University of Pavia, Pavia, Italy. [4]DIISM-SAILAB, University of Siena, Siena, Italy. [5]National Institute for the Infectious Diseases "L. Spallanzani", Rome, Italy. [6]Stem Cell Transplant and Cellular Therapy Unit, University Hospital, Siena, Italy. [7]Istituto di Tecnologie Biomediche – Consiglio Nazionale delle Ricerche, Segrate, MI, Italy. [8]Genetica Medica, Azienda Ospedaliero-Universitaria Senese, Siena, Italy. [9]Department of Experimental Medicine, Sapienza University of Rome, Rome, Italy.

Received: 1 June 2021 Accepted: 3 August 2021
Published online: 16 August 2021

References

1. Tang N, Li D, Wang X, et al. Abnormal coagulation parameters are associated with poor prognosis in patients with novel coronavirus pneumonia. J Thromb Haemost. 2020. https://doi.org/10.1111/jth.14768.
2. Fallerini C, Daga S, Mantovani S, 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. 2021;2(10):e67569. https://doi.org/10.7554/eLife.67569.
3. Severe Covid-19 GWAS Group, Ellinghaus D, Degenhardt F, et al. Genome-wide Association Study of Severe Covid-19 with Respiratory Failure. N Engl J Med. 2020;383(16):1522–34. https://doi.org/10.1056/NEJMoa2020283.
4. Baldassarri M, Picchiotti N, Fava F, et al. Shorter androgen receptor polyQ alleles protect against life-threatening COVID-19 disease in European males. EBioMedicine. 2021;65:103246. https://doi.org/10.1016/j.ebiom.2021.103246.
5. Blann AD, Nadar SK, Lip GYH, et al. The adhesion molecule P-selectin and cardiovascular disease. Eur Heart J. 2003;24:2166–79.
6. Merten M, Thiagarajan P. P-selectin in arterial thrombosis. Z Kardiol. 2004;93(11):855–63. https://doi.org/10.1007/s00392-004-0146-5.
7. Bongiovanni D, Klug M, Lazareva O, et al. SARS-CoV-2 infection is associated with a pro-thrombotic platelet phenotype. Cell Death Dis. 2021;12(1):50. https://doi.org/10.1038/s41419-020-03333-9.
8. Manne BK, Denorme F, Middleton EA, et al. Platelet gene expression and function in patients with COVID-19. Blood. 2020;136(11):1317–29. https://doi.org/10.1182/blood.2020007214.
9. Ay C, Jungbauer LV, Kaider A, et al. P-selectin gene haplotypes modulate soluble P-selectin concentrations and contribute to the risk of venous thromboembolism. Thromb Haemost. 2008;99(5):899–904. https://doi.org/10.1160/TH07-11-0672.
10. Tregouet DA, Barbaux S, Escolano S, et al. Specific haplotypes of the P-selectin gene are associated with myocardial infarction. Hum Mol Genet. 2002;11(17):2015–23. https://doi.org/10.1093/hmg/11.17.2015.
11. Zhang S, Liu Y, Wang X, et al. SARS-CoV-2 binds platelet ACE2 to enhance thrombosis in COVID-19. J Hematol Oncol. 2020;13:120. https://doi.org/10.1186/s13045-020-00954-7.
12. Karolczak K, Konieczna L, Kostka T, et al. Testosterone and dihydrotestosterone reduce platelet activation and reactivity in older men and women. Aging (Albany NY). 2018;10(5):902–29. https://doi.org/10.18632/aging.101438.
13. Agrati C, Bordoni V, Sacchi A, et al. Elevated P-selectin in severe Covid-19: considerations for therapeutic options. Mediterr J Hematol Infect Dis. 2021;13(1):e2021016. https://doi.org/10.4084/MJHID.2021.016.

Publisher's Note

8. Computational prediction of CNVs from WES of COVID-19 infected patients

In the previous chapters, we showed that both polymorphisms and rare variants are involved in COVID-19 severity. However, SNVs are not the only type of variation that can be detected with WES experiments, as we introduced in chapter 1. Therefore, we started evaluating the potential impact of a different type of variation, e.g., copy number variations, in predisposing to COVID-19 disease. Many challenges arise when dealing with computationally predicted CNVs from WES data, as described in chapter 1 section 1.4.4. These shortcomings associated with CNVs detection from WES adds to the complexity of the modelling task, i.e., predicting COVID-19 severity from genetic data. The results presented in this chapter represents a first, preliminary, attempt to study the potential association of CNVs with COVID-19 severity.

8.1 Results of the computational algorithms show striking variation in the length and number of CNVs predicted by the different programs

Results of the computational prediction demonstrated a wide range in both CNV counts and size when using different bioinformatic tools (Figure 7). CoNIFER tends to be more specific and thus detected fewer events of interest (median 11, range 1-1303) while ExomeDepth errs on the side of sensitivity and returned more CNVs per sample (median 392, range 18-695). Also, the size of the predicted CNVs differed between the tools, being 49.94 kbp (median length) for CoNIFER (range 454 bp to 23.089 Mbp) and 1.49 kbp (median length) for ExomeDepth (range 60 bp to 45.4 Mbp).

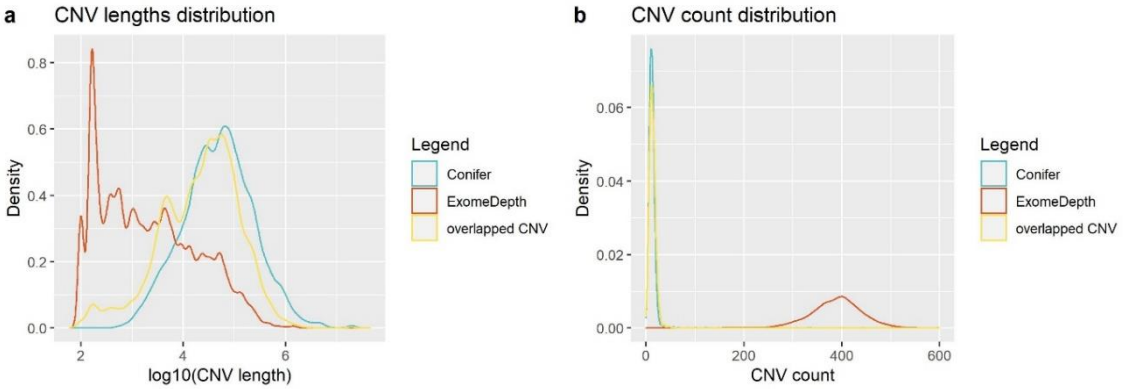


Figure 7. Count and length distributions of CNVs predicted by ExomeDepth (red) and CoNIFER (light blue). In yellow the distributions of overlapped CNVs between the two tools.

8.2 Results of CNVs detected by both CoNIFER and ExomeDepth

As it is essential to filter out results that are unlikely to be true/relevant from our analysis, we selected only the most reliable CNVs, i.e., those predicted by both tools. Since CoNIFER detected fewer and longer events, it was considered as the limiting factor in the contrast. A total of 24850 CNVs were in common between the two tools when the CNV detected by CoNIFER overlapped at least 50% the CNVs predicted by ExomeDepth (-F option). The percentage of overlap was on average equal to 68%. By looking at the reverse contrast, i.e., when imposing that at least 50% of CNV predicted by CoNIFER overlapped with ExomeDepth CNV (-f option), a total of 16130 were found in common (Figure 8). The union of these two sets of overlapped calls was considered for further analyses. 38.3% of events were found in both contrasts as reported in Figure 8. The vast majority of overlapped CNVs derived from the -F contrast as expected. Finally, taking together the results of the two

comparisons, the percentage of overlap among the two tools increases from 68% to 81% (29631 overlapped CNVs out of 36492 detected initially by CoNIFER).

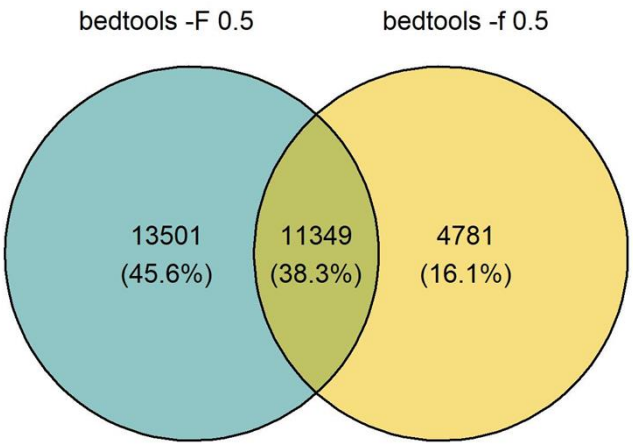


Figure 8. Overlapped CNVs between CoNIFER and ExomeDepth. We imposed 50% of non-reciprocal overlap between the CNVs detected by the two tools. In light blue are reported the overlapped CNVs obtained from the first contrast (-F) and in yellow the ones deriving from the first contrast (-F) and in yellow the ones deriving from the second contrast (-f). 38.3% of intersections were found in both contrasts.

8.3 Results of the LASSO logistic regression: CNVs and COVID-19 severity

The LASSO logistic regression model was fitted on the cohort using as input features the Boolean representation of CNVs described in Chapter 2. In the chosen Boolean representation, the gene was set to 1 if it presented any copy number alteration, and 0 otherwise. Cases were defined as deceased or patients needing endotracheal intubation or CPAP/biPAP ventilation or oxygen support only (category

5/4/3/2). As controls, participants were selected if being hospitalized without oxygen support or oligo/asymptomatic not requiring hospitalization (category 1 and 0). After the fitting of the model, the performances were evaluated by looking at the Receiver Operating Characteristic (ROC) curve (Figure 9, Panel C) which provides an Area Under the Curve (AUC) score of 52%, which is not significantly different than random guess. Results were examined with a Chi Square Test. The first 20 features for importance and the relative p-values are reported in Table 4. No significant association was found after correction for multiple testing.

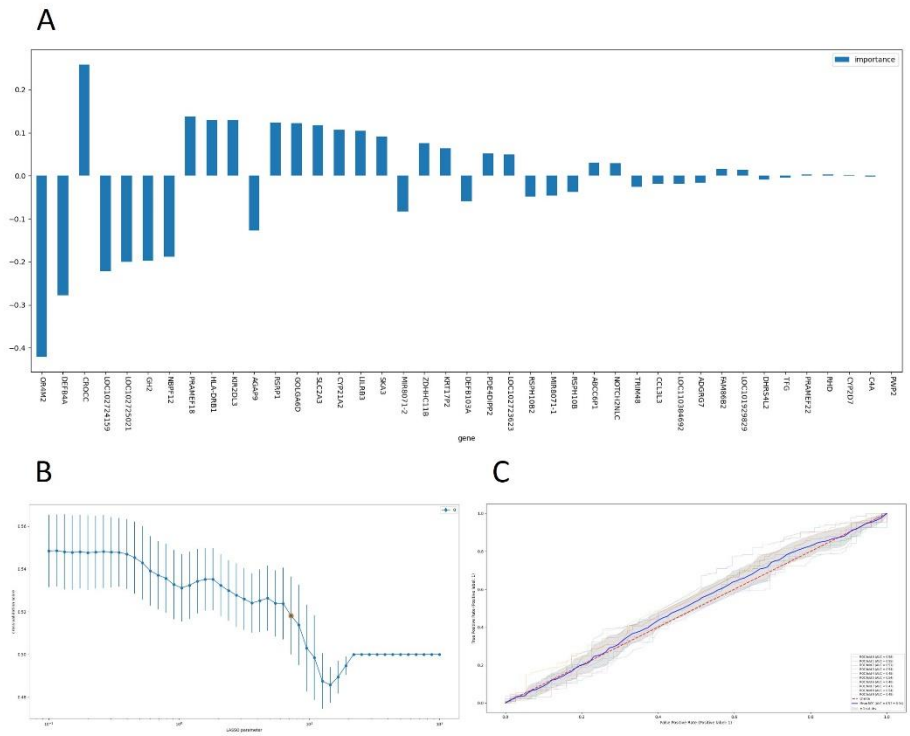


Figure 9. Results of the LASSO logistic regression for the Boolean of CNV. **Panel A.** The histogram of LASSO logistic regression weights represents the importance of each feature for the classification task. **Panel B.** Cross validation ROC-AUC score for the grid of LASSO regularization parameters; the optimal regularization parameter is chosen by selecting the one with highest cross-validation score (red point). **Panel C.** ROC curve for the 10 folds of the cross-validation.

Table 4. First 20 features for importance with the respective p-value in the LASSO logistic regression of the Boolean of CNV. The total number of features tested was 7153.

Feature	P-value
GH2	0.000743
DEFB4A	0.001752
CSH2	0.002236
CSH1	0.002654
ZNF705B	0.00267
OR4M2	0.003571
LOC642846	0.004747
DEFB103A	0.005532
DEFB103B	0.005532
LINC00115	0.006255
FAM87B	0.006255
CROCC	0.006313
FAM66E	0.006696
LOC102725021	0.01201
NBPF12	0.012093
ARHGAP11B	0.013714
MIR3690	0.016334
FRMPD2B	0.016861
LOC102724159	0.017127
PWP2	0.017127

Subsequently, we repeated the same analysis but selecting only those genes subjected to dosage sensitivity. The list of genes was downloaded from https://dosage.clinicalgenome.org/. Results of the LASSO logistic regression are reported in Figure 10. The performance of the model remained low, as the ROC curve for the 10 folds of the cross-validation provided an Area Under the Curve (AUC) score of 51% (Figure 10, Panel C).

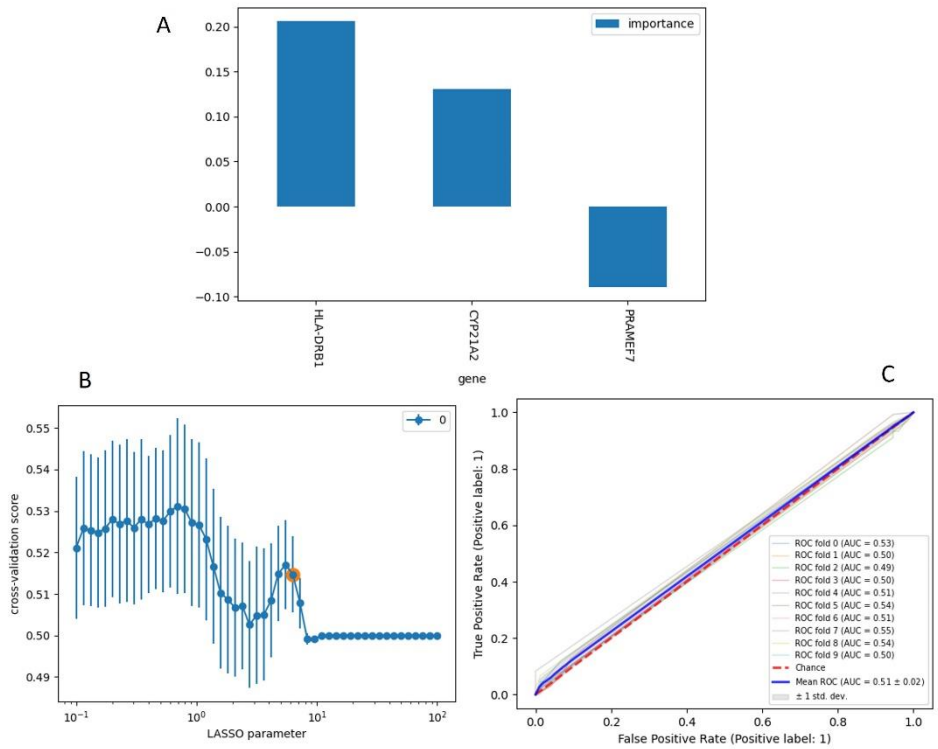


Figure 10. Results of the LASSO logistic regression for the Boolean of CNV filtered by Dosage-sensitive genes. **Panel A.** The histogram of LASSO logistic regression weights represents the importance of each feature for the classification task. **Panel B.** Cross validation ROC-AUC score for the grid of LASSO regularization parameters; the optimal regularization parameter is chosen by selecting the one with highest cross-validation score (red point). **Panel C.** ROC curve for the 10 folds of the cross-validation.

Eventually, we tested deletions and duplications separately using as input features of the model the Boolean representation of deletions and the Boolean representation of duplications, respectively. Results for deletions are reported in Figure 8 and for duplications in Figure 11. Also in these analyses we did not obtain good performance in predicting COVID-19 outcomes.

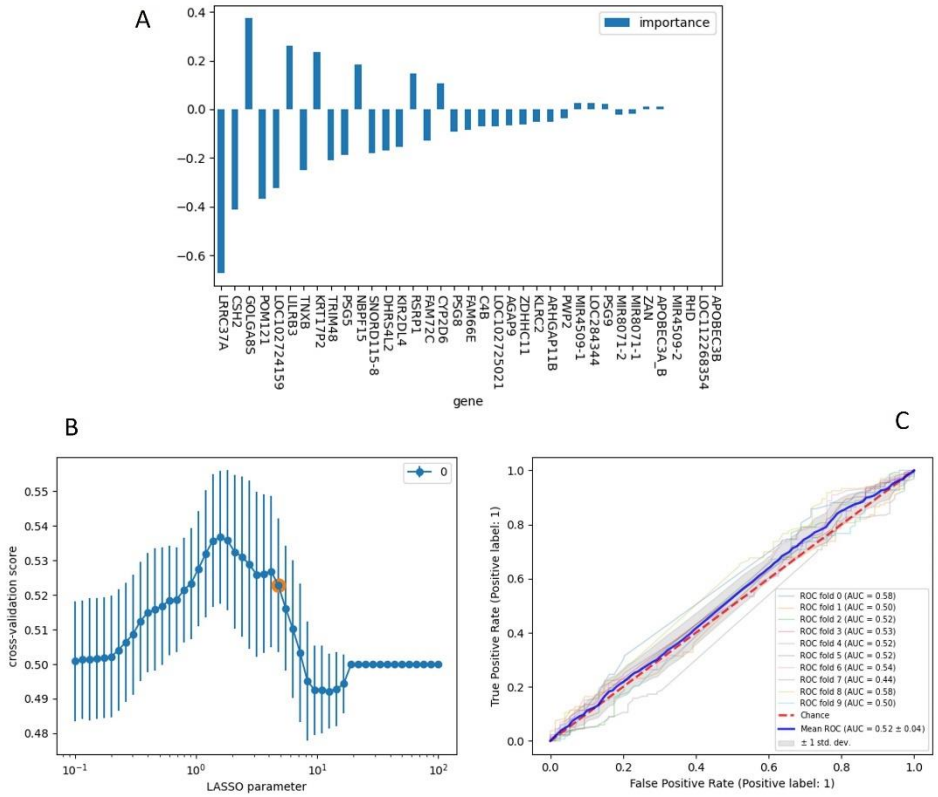


Figure 11. Panel A. The histogram of LASSO logistic regression weights represents the importance of each feature for the classification task. **Panel B.** Cross validation ROC-AUC score for the grid of LASSO regularization parameters; the optimal regularization parameter is chosen by selecting the one with highest cross-validation score (red point). **Panel C.** ROC curve for the 10 folds of the cross-validation.

9. Conclusive remarks and future perspectives

COVID-19 is a condition with a significantly wide range of clinical presentations: from asymptomatic infected patients to those expressing severe symptoms leading to death. Assuming a relatively low impact of different virus variants on the observed interindividual variability, the remaining clinical variability might likely be associated with age and host genetics, including sex. In line with recent studies [12], [14], [16], [49], [63], we focused our attention on the identification of host genetics factors able to explain COVID-19 severity.

This dissertation describes the results obtained with an approach that combines synthetic representations of genetic data and a machine learning model starting from Whole Exome Sequencing data to investigate genetic variability in COVID-19 infected patients. When at the beginning of 2020 we started to collect COVID-19 positive patients from all over Italy in the context of the GEN-COVID Multicenter study, we began to face the complex nature of COVID-19 infection. We soon realised that host genetics could play an important role in COVID-19 pathogenesis. By looking at the coding variants in the *ACE2* gene, the SARS-CoV-2 receptor for host cell entry, we found a statistically significant higher allelic heterogeneity for *ACE2* in controls compared to cases, with a higher chance to find at least one *ACE2* variant in the cohort of controls compared to the cohort of cases [10]. We therefore suggested that the effect of rare variants, likely summing up to the effect of more frequent ones, could partially account for the inter-individual clinical variability observed.

This initial hypothesis was further explored in a subsequent pilot study [50] where common variants in susceptibility genes seemed to represent the favourable background in which additional host private mutations may determine disease progression. We realized the need for a new method that could combine common and rare variants and, at the same time, extract relevant information from the massive datasets derived from WES experiments. We therefore proposed a new approach to

identify host risk factors predisposing to the disease. The innovation consisted in mapping the genetic variability into a set of informative features, e.g., Boolean representations, to predict the COVID-19 severity using LASSO logistic regression.

The first analysis carried out by exploiting this method on the dataset of COVID-19 was aimed to understand if the differences observed in the outcomes between men and women could be explained by the host genome. Epidemiological studies, in fact, indicate that men and women are similarly infected by SARS-CoV-2, but COVID-19 outcome is less favourable in men. In this study, reported in chapter 5, we identified the first genetic polymorphism predisposing some men to develop a more severe disease, irrespectively of age, by comparing the extreme ends of the cohort (severe vs. oligo-asymptomatic SARS-CoV-2 PCR-positive patients). We demonstrated that the number of polyQ repeats in the androgen receptor (*AR*) gene is a predictor of the COVID-19 outcome as polyQ alleles shorter than 22 repeats in the receptor conferred protection against severe outcome in COVID-19, independently of age. Failure of the endocrine feedback to overcome AR signalling defect by increasing testosterone levels during the infection leads to the fact that polyQ becomes dominant to testosterone levels for the clinical outcome [59]. This first result opens potential of using testosterone as adjuvant therapy for patients with severe COVID-19 having defective androgen signalling, defined in this study as PolyQ ≥ 23 repeats, and inappropriately low levels of circulating androgens. This study shows a successful application of the LASSO logistic regression on the Boolean of poly-amino acids triplet repeats.

Subsequently, we focused our attention on rare genetic variants. In the study reported in chapter 6, we analysed rare variants (MAF<1%) on X chromosome by comparing young males (<60 years) of the extreme phenotypes of the GEN-COVID cohort. LASSO logistic regression on the XL Boolean feature picked up *TLR7* as the most important susceptibility gene. Loss-of-function variants in the X-linked recessive *TLR7* Mendelian form contributed to disease susceptibility in up to 2% of severe COVID-19 [60]. These results were validated by functional gene expression profile demonstrating a reduction in *TLR7*-related gene expression in cases compared

to controls, underling an impairment in type I and II IFN responses. We therefore confirmed the role of *TLR7* in COVID-19 susceptibility in young males, previously reported by van der Made *et* colleagues, extending the results in a larger cohort. These findings were further validated by other research groups [56].

In chapter 7, we identified a common polymorphism, Asp603Asn in *SELP*, associated with severity and thromboembolism, leading to life-threatening disease. This result was obtained using LASSO logistic regression on the Boolean representation of homozygous common bi-allelic polymorphism of autosomal genes in males. In this study we showed that predisposition to thromboembolism increases if the protective effect of testosterone is lost either by age or because of additional genetic factors such as polyQ ≥23 in the *AR* gene [62]. This knowledge provides a rationale for repurposing anti P-selectin monoclonal antibodies as personalized adjuvant therapy in men affected by COVID-19.

All these results together show that this novel synthetic approach was effective to characterize both common and rare variants as potential contributors to the severe phenotypes, providing knowledge for potential patients' treatment.

To evaluate if other type of variation, e.g., Copy Number Variants, could account for a part of COVID-19 heritability, we built a Boolean representation of CNV to be tested with the same strategy. Limitations associated with CNVs detection from WES increases the complexity of the modelling task, i.e., predicting COVID-19 severity from genetic data. In these preliminary results presented in chapter 8, we did not find any association between CNVs, computationally predicted by two independent tools, and COVID-19 severity. However, further studies are necessary to assess their potential contribution to COVID-19 outcomes.

In conclusion, the approaches presented in this thesis allowed to identify several genetic factors responsible for interindividual variability in the response to SARS-COV2 infections. The natural evolution of the work presented here is the development of a comprehensive model, that combines the different representations of genetic variability into a unified framework. A first attempt in this direction has

recently been proposed in the context of the GEN-COVID consortium [65] and in collaboration with international cohorts contributing to the WES/WGS working group within the HGI (https://www.covid19hg.org/projects/). In the mentioned study, we propose an Integrated PolyGenic Score (IPGS) that includes information regarding the variants at different frequencies, from ultra-rare to common. The input features of the model are the gene-based Boolean features presented in section 2.5.1. Severity predictions considering IPGS as an input feature were shown to outperform predictions not considering the genetic information [65].

This novel approach can significantly improve our ability to estimate the contribution of genetic factors to the risk of suffering a severe form of COVID-19 and can help to understand the potential implications for clinical and public health responses. Moreover, besides the relevance for the current pandemic, the methods presented could help us to understand the role of genetics in other complex diseases.

Bibliography

[1] Z. Wu and J. M. McGoogan, “Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China,” *JAMA*, vol. 323, no. 13, Apr. 2020, doi: 10.1001/jama.2020.2648.

[2] H. Deng, X. Yan, and L. Yuan, “Human genetic basis of coronavirus disease 2019,” *Signal Transduction and Targeted Therapy*, vol. 6, no. 1, p. 344, Dec. 2021, doi: 10.1038/s41392-021-00736-8.

[3] WHO. Coronavirus (COVID-19) Dashboard., “https:// covid19.who.int/ (accessed October 15, 2021)”.

[4] N. He, “Rapid evolution of the COVID-19 pandemic calls for a unified public health response,” *BioScience Trends*, vol. 15, no. 4, Aug. 2021, doi: 10.5582/bst.2021.01261.

[5] C. Fernández-de-las-Peñas, “Long COVID: current definition,” *Infection*, Sep. 2021, doi: 10.1007/s15010-021-01696-5.

[6] M. O’Driscoll *et al.*, “Age-specific mortality and immunity patterns of SARS-CoV-2,” *Nature*, vol. 590, no. 7844, Feb. 2021, doi: 10.1038/s41586-020-2918-0.

[7] E. P. Scully, J. Haverfield, R. L. Ursin, C. Tannenbaum, and S. L. Klein, “Considering how biological sex impacts immune responses and COVID-19 outcomes,” *Nature Reviews Immunology*, vol. 20, no. 7, Jul. 2020, doi: 10.1038/s41577-020-0348-8.

[8] J. Y. Ko *et al.*, “Risk Factors for Coronavirus Disease 2019 (COVID-19)–Associated Hospitalization: COVID-19–Associated Hospitalization Surveillance Network and Behavioral Risk Factor Surveillance System,” *Clinical Infectious Diseases*, vol. 72, no. 11, Jun. 2021, doi: 10.1093/cid/ciaa1419.

[9] G. Onder, G. Rezza, and S. Brusaferro, “Case-Fatality Rate and Characteristics of Patients Dying in Relation to COVID-19 in Italy,” *JAMA*, Mar. 2020, doi: 10.1001/jama.2020.4683.

[10] E. Benetti *et al.*, “ACE2 gene variants may underlie interindividual variability and susceptibility to COVID-19 in the Italian population,” *European Journal of Human Genetics*, vol. 28, no. 11, pp. 1602–1614, Nov. 2020, doi: 10.1038/s41431-020-0691-z.

[11] D. Yesudhas, A. Srivastava, and M. M. Gromiha, "COVID-19 outbreak: history, mechanism, transmission, structural studies and therapeutics," *Infection*, vol. 49, no. 2, Apr. 2021, doi: 10.1007/s15010-020-01516-2.

[12] COVID-19 Host Genetics Initiative, "Mapping the human genetic architecture of COVID-19," *Nature*, 2021, doi: 10.1038/s41586-021-03767-x.

[13] M. G. P. van der Wijst *et al.*, "Type I interferon autoantibodies are associated with systemic immune alterations in patients with COVID-19," *Science Translational Medicine*, vol. 13, no. 612, Sep. 2021, doi: 10.1126/scitranslmed.abh2624.

[14] P. Bastard *et al.*, "Autoantibodies against type I IFNs in patients with life-threatening COVID-19," *Science*, vol. 370, no. 6515, Oct. 2020, doi: 10.1126/science.abd4585.

[15] The Severe Covid-19 GWAS Group, "Genomewide Association Study of Severe Covid-19 with Respiratory Failure," *New England Journal of Medicine*, vol. 383, no. 16, Oct. 2020, doi: 10.1056/NEJMoa2020283.

[16] E. Pairo-Castineira *et al.*, "Genetic mechanisms of critical illness in COVID-19," *Nature*, vol. 591, no. 7848, Mar. 2021, doi: 10.1038/s41586-020-03065-y.

[17] P. Suwinski, C. K. Ong, M. H. T. Ling, Y. M. Poh, A. M. Khan, and H. S. Ong, "Advancing personalized medicine through the application of whole exome sequencing and big data analytics," *Frontiers in Genetics*, vol. 10, no. FEB. Frontiers Media S.A., 2019. doi: 10.3389/fgene.2019.00049.

[18] Y. Guo, Y. Dai, H. Yu, S. Zhao, D. C. Samuels, and Y. Shyr, "Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis," *Genomics*, vol. 109, no. 2, pp. 83–90, Mar. 2017, doi: 10.1016/j.ygeno.2017.01.005.

[19] A. Sathyanarayanan, S. Manda, M. Poojary, and S. H. Nagaraj, "Exome sequencing data analysis," in *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, vol. 1–3, Elsevier, 2018, pp. 164–175. doi: 10.1016/B978-0-12-809633-8.20094-0.

[20] G. A. Rouleau, J. P. Ross, and P. A. Dion, "Exome sequencing in genetic disease: Recent advances and considerations," *F1000Research*, vol. 9. F1000 Research Ltd, 2020. doi: 10.12688/f1000research.19444.1.

[21] The 1000 Genomes Project Consortium, "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, no. 7319, Oct. 2010, doi: 10.1038/nature09534.

[22] M. J. Bamshad *et al.*, “Exome sequencing as a tool for Mendelian disease gene discovery,” *Nature Reviews Genetics*, vol. 12, no. 11. pp. 745–755, Nov. 2011. doi: 10.1038/nrg3031.

[23] H. Sun and G. Yu, “New insights into the pathogenicity of non-synonymous variants through multi-level analysis,” *Scientific Reports*, vol. 9, no. 1, Dec. 2019, doi: 10.1038/s41598-018-38189-9.

[24] D. S. Marchuk *et al.*, “Increasing the diagnostic yield of exome sequencing by copy number variant analysis,” *PLoS ONE*, vol. 13, no. 12, Dec. 2018, doi: 10.1371/journal.pone.0209185.

[25] L. Feuk, A. R. Carson, and S. W. Scherer, “Structural variation in the human genome,” *Nature Reviews Genetics*, vol. 7, no. 2. pp. 85–97, Feb. 2006. doi: 10.1038/nrg1767.

[26] C. Alkan, B. P. Coe, and E. E. Eichler, “Genome structural variation discovery and genotyping,” *Nature Reviews Genetics*, vol. 12, no. 5. pp. 363–376, May 2011. doi: 10.1038/nrg2958.

[27] O. Pös *et al.*, “Copy number variation: Characteristics, evolutionary and pathological aspects,” *Biomedical Journal*, Feb. 2021, doi: 10.1016/j.bj.2021.02.003.

[28] J. R. Lupski and P. Stankiewicz, “Genomic disorders: Molecular mechanisms for rearrangements and conveyed phenotypes,” *PLoS Genetics*, vol. 1, no. 6. pp. 0627–0633, 2005. doi: 10.1371/journal.pgen.0010049.

[29] R. Beroukhim *et al.*, “The landscape of somatic copy-number alteration across human cancers,” *Nature*, vol. 463, no. 7283, Feb. 2010, doi: 10.1038/nature08822.

[30] A. C. Fahed, B. D. Gelb, J. G. Seidman, and C. E. Seidman, “Genetics of Congenital Heart Disease,” *Circulation Research*, vol. 112, no. 4, Feb. 2013, doi: 10.1161/CIRCRESAHA.112.300853.

[31] E. Bacchelli *et al.*, “An integrated analysis of rare CNV and exome variation in Autism Spectrum Disorder using the Infinium PsychArray,” *Scientific Reports*, vol. 10, no. 1, Dec. 2020, doi: 10.1038/s41598-020-59922-3.

[32] E. Gonzalez, “The Influence of CCL3L1 Gene-Containing Segmental Duplications on HIV-1/AIDS Susceptibility,” *Science*, vol. 307, no. 5714, Mar. 2005, doi: 10.1126/science.1101160.

[33] F. Zhang, W. Gu, M. E. Hurles, and J. R. Lupski, “Copy number variation in human health, disease, and evolution,” *Annual Review of Genomics and*

Human Genetics, vol. 10. pp. 451–481, Sep. 2009. doi: 10.1146/annurev.genom.9.081307.164217.

[34] D. R. Schrider *et al.*, "Gene Copy-Number Polymorphism Caused by Retrotransposition in Humans," *PLoS Genetics*, vol. 9, no. 1, Jan. 2013, doi: 10.1371/journal.pgen.1003242.

[35] W. Li and M. Olivier, "Current analysis platforms and methods for detecting copy number variation," *Physiol Genomics*, vol. 45, pp. 1–16, 2013, doi: 10.1152/physiolgenomics.00082.2012.-Copy.

[36] J. P. Schouten, "Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification," *Nucleic Acids Research*, vol. 30, no. 12, Jun. 2002, doi: 10.1093/nar/gnf056.

[37] L. Zhang, W. Bai, N. Yuan, and Z. Du, "Comprehensively benchmarking applications for detecting copy number variation," *PLoS Computational Biology*, vol. 15, no. 5, 2019, doi: 10.1371/journal.pcbi.1007069.

[38] S. M. Teo, Y. Pawitan, C. S. Ku, K. S. Chia, and A. Salim, "Statistical challenges associated with detecting copy number variations with next-generation sequencing," *Bioinformatics*, vol. 28, no. 21. pp. 2711–2718, Nov. 2012. doi: 10.1093/bioinformatics/bts535.

[39] J. Y. Hehir-Kwa, R. Pfundt, and J. A. Veltman, "Exome sequencing and whole genome sequencing for the detection of copy number variation," *Expert Review of Molecular Diagnostics*, vol. 15, no. 8, Aug. 2015, doi: 10.1586/14737159.2015.1053467.

[40] F. Zare, M. Dow, N. Monteleone, A. Hosny, and S. Nabavi, "An evaluation of copy number variation detection tools for cancer using whole exome sequencing data," *BMC Bioinformatics*, vol. 18, no. 1, May 2017, doi: 10.1186/s12859-017-1705-x.

[41] M. Zhao, Q. Wang, Q. Wang, P. Jia, and Z. Zhao, "Computational tools for copy number variation (CNV) detection using next-generation sequencing data: Features and perspectives," *BMC Bioinformatics*, vol. 14, no. SUPPL11, Sep. 2013, doi: 10.1186/1471-2105-14-S11-S1.

[42] A. Magi, L. Tattini, T. Pippucci, F. Torricelli, and M. Benelli, "Read Count approach for DNA copy number variants detection." [Online]. Available: http://bioinformatics.oxfordjournals.org/

[43] S. Välipakka *et al.*, "Improving Copy Number Variant Detection from Sequencing Data with a Combination of Programs and a Predictive Model," *Journal of Molecular Diagnostics*, vol. 22, no. 1, pp. 40–49, Jan. 2020, doi: 10.1016/j.jmoldx.2019.08.009.

[44] V. Plagnol *et al.*, “A robust model for read count data in exome sequencing experiments and implications for copy number variant calling,” *Bioinformatics*, vol. 28, no. 21, pp. 2747–2754, Nov. 2012, doi: 10.1093/bioinformatics/bts526.

[45] N. Krumm *et al.*, “Copy number variation detection and genotyping from exome sequence data,” *Genome Research*, vol. 22, no. 8, pp. 1525–1532, Aug. 2012, doi: 10.1101/gr.138115.112.

[46] Y. Benjamini and T. P. Speed, “Summarizing and correcting the GC content bias in high-throughput sequencing,” *Nucleic Acids Research*, vol. 40, no. 10, May 2012, doi: 10.1093/nar/gks001.

[47] A. Abyzov, A. E. Urban, M. Snyder, and M. Gerstein, “CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing,” *Genome Research*, vol. 21, no. 6, Jun. 2011, doi: 10.1101/gr.114876.110.

[48] S. Yoon, Z. Xuan, V. Makarov, K. Ye, and J. Sebat, “Sensitive and accurate detection of copy number variants using read depth of coverage,” *Genome Research*, vol. 19, no. 9, Sep. 2009, doi: 10.1101/gr.092981.109.

[49] C. I. van der Made *et al.*, “Presence of Genetic Variants Among Young Men With Severe COVID-19,” *JAMA*, vol. 324, no. 7, Aug. 2020, doi: 10.1001/jama.2020.13719.

[50] E. Benetti *et al.*, “Clinical and molecular characterization of COVID-19 hospitalized patients,” *PLoS ONE*, vol. 15, no. 11 November, Nov. 2020, doi: 10.1371/journal.pone.0242534.

[51] T. G. Dietterich, “Machine Learning for Sequential Data: A Review,” 2002. doi: 10.1007/3-540-70659-3_2.

[52] I. H. Sarker, “Machine Learning: Algorithms, Real-World Applications and Research Directions,” *SN Computer Science*, vol. 2, no. 3, May 2021, doi: 10.1007/s42979-021-00592-x.

[53] J. A. M. Sidey-Gibbons and C. J. Sidey-Gibbons, “Machine learning in medicine: a practical introduction,” *BMC Medical Research Methodology*, vol. 19, no. 1, Dec. 2019, doi: 10.1186/s12874-019-0681-4.

[54] Robert Tibshirani, “Regression Shrinkage and Selection via the Lasso,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, pp. 267–288, 1996, Accessed: Sep. 27, 2021. [Online]. Available: https://www.jstor.org/stable/2346178

[55] COVID-19 Therapeutic Trial Synopsis, "novel Coronavirus," *WHO R&D Blueprint*, 2020.

[56] H. Li *et al.*, "The Sequence Alignment/Map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, Aug. 2009, doi: 10.1093/bioinformatics/btp352.

[57] A. R. Quinlan and I. M. Hall, "BEDTools: a flexible suite of utilities for comparing genomic features," *Bioinformatics*, vol. 26, no. 6, Mar. 2010, doi: 10.1093/bioinformatics/btq033.

[58] V. Geoffroy *et al.*, "AnnotSV: an integrated tool for structural variations annotation," *Bioinformatics*, vol. 34, no. 20, Oct. 2018, doi: 10.1093/bioinformatics/bty304.

[59] M. Baldassarri *et al.*, "Shorter androgen receptor polyQ alleles protect against life-threatening COVID-19 disease in European males," *EBioMedicine*, vol. 65, Mar. 2021, doi: 10.1016/j.ebiom.2021.103246.

[60] C. Fallerini *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, Mar. 2021, doi: 10.7554/eLife.67569.

[61] S. Zhang *et al.*, "SARS-CoV-2 binds platelet ACE2 to enhance thrombosis in COVID-19," *Journal of Hematology & Oncology*, vol. 13, no. 1, Dec. 2020, doi: 10.1186/s13045-020-00954-7.

[62] C. Fallerini *et al.*, "SELP Asp603Asn and severe thrombosis in COVID-19 males," *Journal of Hematology & Oncology*, vol. 14, no. 1, p. 123, Dec. 2021, doi: 10.1186/s13045-021-01136-9.

[63] Q. Zhang *et al.*, "Inborn errors of type I IFN immunity in patients with life-threatening COVID-19," *Science*, vol. 370, no. 6515, Oct. 2020, doi: 10.1126/science.abd4570.

[64] T. Asano *et al.*, "X-linked recessive TLR7 deficiency in ~1% of men under 60 years old with life-threatening COVID-19," *Science Immunology*, vol. 6, no. 62, Aug. 2021, doi: 10.1126/sciimmunol.abl4348.

[65] C. Fallerini *et al.*, "Common, low-frequency, rare, and ultra-rare coding variants contribute to COVID-19 severity," *medRxiv*, 2021.