Common, low-frequency, rare, and ultra-rare coding variants contribute to COVID-19 severity

Chiara Fallerini1,2 · Nicola Picchiotti3,4 · Margherita Baldassarri1,2 · Kristina Zgroj1 · Sergio Daga1,2 · Francesca Fava1,2,5 · Elisa Benetti1 · Sara Amitrano5 · Mirella Bruttini1,2 · Maria Palmieri1,2 · Susanna Croci1,2 · Mirjam Lista1,2 · Giada Beligni1,2 · Floriana Valentino1,2 · Ilaria Meloni1,2 · Marco Tanfoni3 · Francesca Minnai6 · Francesca Colombo6 · Enrico Cabri7 · Maddalena Fratelli7 · Chiara Gabbi8 · Stefania Mantovani9 · Elisa Frullanti1,2 · Marco Gori3,10 · Francis P. Crawley11 · Guillaume Butler-Laporte12,13 · Brent Richards12,14,15 · Hugo Zeberg16 · Miklos Lipcsey17,18 · Michael Hultström17,19 · Kerstin U. Ludwig20 · Eva C. Schulte21,22,23 · Erola Pairo-Castineira24,25 · John Kenneth Baillie24,25,26 · Axel Schmidt20 · Robert Frithiof17 · WES/WGS Working Group Within the HGI · GenOMICC Consortium · GEN-COVID Multicenter Study · Francesca Mari1,2,5 · Alessandra Renieri1,2,5,27 · Simone Furini1

Received: 14 September 2021 / Accepted: 26 October 2021 / Published online: 10 December 2021
© The Author(s) 2021

Abstract
The combined impact of common and rare exonic variants in COVID-19 host genetics is currently insufficiently understood. Here, common and rare variants from whole-exome sequencing data of about 4000 SARS-CoV-2-positive individuals were used to define an interpretable machine-learning model for predicting COVID-19 severity. First, variants were converted into separate sets of Boolean features, depending on the absence or the presence of variants in each gene. An ensemble of LASSO logistic regression models was used to identify the most informative Boolean features with respect to the genetic bases of severity. The Boolean features selected by these logistic models were combined into an Integrated PolyGenic Score that offers a synthetic and interpretable index for describing the contribution of host genetics in COVID-19 severity, as demonstrated through testing in several independent cohorts. Selected features belong to ultra-rare, rare, low-frequency, and common variants, including those in linkage disequilibrium with known GWAS loci. Noteworthily, around one quarter of the selected genes are sex-specific. Pathway analysis of the selected genes associated with COVID-19 severity reflected the multi-organ nature of the disease. The proposed model might provide useful information for developing diagnostics and therapeutics, while also being able to guide bedside disease management.

Introduction
For almost 2 years, COVID-19 has demonstrated itself to be a disease having a broad spectrum of clinical presentations: from asymptomatic patients to those with severe symptoms leading to death or persistent disease (“long COVID”) (Livingston and Bucher 2020; Chen et al. 2019; Zhang et al. 2020a). While developing vaccination programmes and other preventive measures to significantly dampen infection transmission and reduce disease expression, a much deeper and more precise understanding of the interplay between SARS-CoV-2 and host genetics is required to support the development of treatments for new virus variants as they arise. Furthermore, advances in modelling the interplay between SARS-CoV-2 and host genetics hold significant promise for addressing other complex diseases. In this study,
we demonstrate the value of genetic modelling with its direct translatability into drug development and clinical care in the context of a severe public health crisis.

The identification of host genetic factors modifying disease susceptibility and/or disease severity has the potential to reveal the biological basis of disease susceptibility and outcome as well as to subsequently contribute to treatment amelioration (Elhabyan et al. 2020). From a scientific point of view, COVID-19 represents a particularly interesting and accessible complex disorder for modeling host genetic data because the environmental factor (SARS-CoV-2) can be readily identified by a PCR-based swab test. The still moderate viral genome variability has thus far been shown to have relatively low impact on disease severity (Islam et al. 2020) where currently age, sex, and comorbidities are the major factors predicting disease susceptibility and outcome (Li et al. 2021). While these factors certainly have significant value for prediction, they provide limited insights into disease pathophysiology and are of limited relevance for drug development.

Common variants in the human genome affecting the susceptibility to SARS-CoV-2 infections and COVID-19 severity have been successfully identified by Genome-Wide Association Studies (GWASs) (Severe Covid-19 GWAS Group 2020; Pairo-Castinea et al. 2020; COVID-19 Host Genetics Initiative et al. 2021). However, these variants only explain a small fraction of trait variability and, as it is well documented, GWASs are difficult to interpret because they often associate non-coding variants with phenotype; therefore, the relevant genes need to be pinpointed by deeper follow-up analyses. In contrast, next-generation sequencing-based studies have identified variants in a few genes-related to innate immunity which can solely underlie rare severe forms of COVID-19 (Zhang et al. 2020b; Van der Made et al. 2020; Fallerini et al. 2021a; Solanich et al. 2021). In these rare affected families, the predictivity is high as the susceptibility for severe COVID-19 follows Mendelian inheritance patterns. However, these patients represent only a small proportion of those severely affected by COVID-19. Taken together, genetic findings can currently only explain a limited proportion of COVID-19 susceptibility and severity, in spite of the relatively high predicted heritability of COVID-19 and COVID-19 symptoms (Williams et al. 2020). A better and more holistic understanding of host genetics could support the development of more specific, or even targeted drugs and treatment interventions leading to less morbidity and mortality.

The Italian GEN-COVID Multicenter Study collected more than 2000 biospecimens and clinical data from SARS-CoV-2-positive individuals (Daga et al. 2021), and whole-exome sequencing (WES) analysis contributed to the identification of rare variants (Fallerini et al. 2021a) and common polymorphisms (Baldassarri et al. 2021a; Croci et al. 2021; Fallerini et al. 2021b) associated with COVID-19 severity. 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 (Benetti et al. 2020a). We then proposed a new post-Mendelian model for a genetic characterization of the disorder (Picchiotti et al. 2021) based on an adapted Polygenic Risk Score (PRS) (Mars et al. 2020), called Integrated PolyGenic Score (IPGS). This allowed us to reach a more precise disease severity prediction than that based on sex and age alone. In this article, we substantially improve this post-Mendelian model to include ultra-rare and low-frequency variants while also demonstrating that IPGS significantly contributes to predictivity in combination with—as well as alongside—age and sex, and is able to extract patient-specific genes. The IPGS predictivity was also sustained in three independent European cohorts of the WES/Whole-Genome Sequencing study working group within the COVID-19 Host Genetics Initiative (COVID-19 Host Genetics Initiative 2021).

Materials and methods

Contributing cohorts

Five different cohorts (from Germany, Italy, Quebec, Sweden, UK) contributed to this study as described in Supplementary Table 1. For multi ancestry cohorts (Quebec and UK), the subpopulation of European Ancestry was included in the study. Institutional Review Board approval was obtained for each study (Supplementary Table 1).

Phenotype definitions

The training of the model proposed for predicting the severity of COVID-19 requires as inputs the exome variants, age, sex, and COVID-19 severity assessed using a modified version of the WHO COVID-19 Outcome Scale (COVID-19 Therapeutic Trial Synopsis 2020) as coded into the following six classifications: (1) death; (2) hospitalized receiving invasive mechanical ventilation; (3) hospitalized, receiving continuous positive airway pressure (CPAP) or bilevel positive airway pressure (BiPAP) ventilation; (4) hospitalized, receiving low-flow supplemental oxygen; (5) hospitalized, not receiving supplemental oxygen; and (6) not hospitalized. The aim of the model is to predict a binary classification of patients into mild and severe cases, where a patient is considered severe if hospitalized and receiving any form of respiratory support (WHO severity-grading equal to 4 or higher in 8 points classification). The next section describes
how the annotation of exome variants and the selection of patients were performed in the GEN-COVID cohort. Following this, the training and testing of the model are described.

**Massive parallel sequencing**

**GEN-COVID cohort**

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

**Swedish cohort**

Whole-exome sequencing was performed using the Twist Bioscience exome capture probe and was sequenced on the Illumina NovaSeq6000 platform. Data were then analyzed using the McGill Genome Center bioinformatics pipeline (https://doi.org/10.1093/gigascience/giz037) in accordance with GATK best practice guidelines.

**DeCOI Germany**

800–1000 ng of genomic DNA of each individual was fragmented to an average length of 350 bp. Library preparation was performed using the TruSeq DNA PCR-free kit (Illumina, San Diego, CA, USA) according to the manufacturer’s protocol. Whole-genome sequences were obtained as 150 bp paired-end reads on S4 flow cells using the NovaSeq6000 system (Illumina). The intended average sequencing depth was 30X. The DRAGEN pipeline (Illumina, version 3.6.3 or 3.5.7) was used for alignment and joint variant calling was performed with the Glnexus software (version 1.3.2). Individuals with a 20-fold coverage in less than 96% of the protein-coding sequence were removed as well as related individuals to retain only from related pairs. Variant QC was performed using hail (version 0.2.58). European individuals were selected by performing PCA analysis along with the 1000 genomes data. Finally, annotation was performed using Variant Effect Predictor (VEP, version 101). In the present study, only individuals from the BoSCO and CoMRI substudies were included.

**BQC-19**

Whole-genome sequencing at mean coverage of 30x was performed on the Illumina NovaSeq6000 platform, then analyzed using the McGill Genome Center bioinformatics pipeline (https://doi.org/10.1093/gigascience/giz037), in accordance with GATK best practice guidelines.

**GenOMICC/ISARIC4C**

Whole-genome sequencing at mean coverage of 20x was performed on the Illumina NovaSeq6000 platform and then analyzed using the DRAGEN pipeline (software v01.011.269.3.2.22 , hardware v01.011.269). Variants were genotyped with the GATK GenotypeGVCFs tool v4.1.8.1.

**PC analysis**

The genetic ancestry of the patients was estimated using a random forest classifier. SNPs of autosomes with MAF above 10% and in linkage disequilibrium were extracted from the 1000 genome project using BCFTOOLS (Danecek et al. 2021), and intersected with variants from the GEN-COVID cohort. The resulting set of variants was used to compute 6 Principal Components by PLINK (Purcell et al. 2007) using samples from the 1000 genome projects. GEN-COVID samples were projected along the same Principal Components. The Random Forest classifier, as implemented in the R library randomForest (Liaw and Wiener 2002), was trained using samples from 1000 genomes with known ancestry, and then used to predict ancestry for the GEN-COVID cohort. To avoid bias in the analysis due to the different ethnicity, only patients of genetic European ancestry were retained for further analyses.

**Definition of the Boolean features**

Variants were converted into 12 sets of Boolean features to better represent the variability at the gene-level. First, any variant not impacting on the protein sequence was discarded. Then, the remaining variants were classified according to their minor allele frequency (MAF) as reported in gnomAD for the reference population as: ultra-rare, MAF < 0.1%; rare, 0.1% ≤ MAF < 1%; low-frequency, 1% ≤ MAF < 5%; and common, MAF ≥ 5%. Non-Finnish European (NFE) was used as a reference population. SNPs with MAF not available in gnomAD were treated as ultra-rare. INDELs with frequency not available in gnomAD were treated as ultra-rare when present only once in the cohort and otherwise discarded as possible artefacts of sequencing. For the ultra-rare variants, 3 alternative Boolean representations were defined, which are designed to capture the autosomal dominant (AD), autosomal recessive (AR), and X-linked...
analysis), the genes already demonstrated to be associated with severity in functional studies and the list deduced from integrative genomic analyses by Pathak et al. (2021).

Model training

The dataset was divided into a training set and a testing set (90/10), and the entire procedure described in this section was performed using only samples in the training set. A bootstrap approach with 100 iterations was adopted to train the model. At each bootstrap iteration, 90% of the samples were selected (without replication), and the following two steps were performed: (step 1) selection of the most relevant features for each Boolean representations; and (step 2) definition of the weights of the Integrated Polygenic Score (IPGS). After the 100 bootstrap iterations, the information extracted on relevant features and weighting factors are merged to define the final IPGS (step 3). The IPGS is then used, together with age and sex, for training a model that predicts the COVID-19 severity (step 4). These four steps of the training procedure are described in detail in the next subsections. Since the model is based on a combination of rare and common variants, the training procedure should be performed using a dataset with homogeneous ancestry.

Step 1: Features’ selection

The subsets of the most relevant features were identified using logistic regression models with least absolute shrinkage and selection operator (LASSO) regularization. Separate logistic models were trained for each of the 12 sets of Boolean features. The predicted outcome variable for each of these models was a re-classified phenotype adjusted by age and sex. To compute these re-classified phenotypes as adjusted by age and sex, the patients were first divided into males and females. Then, for each sex, an ordinal logistic regression model was fitted using the age to predict the WHO phenotype classification into six grades. The ordinal logistic regression model was chosen as: it imposes a simple monotone relation between input feature and target variable; and it provides easily interpretable thresholds between the predicted classes. The patients with a predicted grading equal to the actual grading were excluded. The remaining patients were divided into two classes depending on whether their actual phenotype was milder or more severe than the one expected for a patient of that age and sex. This procedure has the benefit of isolating patients whose genetic factors are most important for predicting COVID-19 severity. This binary trait, i.e. phenotype more/less severe than expected, was used as the outcome variable for the 12 LASSO logistic models based on the 12 separate Boolean representations. For each LASSO model, the regularization strength was optimized by tenfold cross-validation with 50
equally spaced values in the logarithmic scale in the range $[10^{-2}, 10^1]$. The optimal regularization strength was selected as the one with the best trade-off between the simplicity of the model and the cross-validation score, i.e. as the highest regularization strength providing an average score closer to the highest average score than 0.5 standard deviations. Once the regularization strength was defined, the LASSO model was re-trained using all the samples in that particular bootstrap iteration. The features with non-null coefficients are the ones selected for the next step. In summary, for each bootstrap iteration, this procedure returns 12 lists of features (one for each Boolean representation) that are expected to be the most important features for predicting the phenotype adjusted by age and sex (in that particular bootstrap iteration).

**Step 2: Weights of the Integrated Polygenic Score (IPGS)**

In the previous step, the Boolean representations are considered isolated from each other. The aim of the IPGS is to combine information from different representations (Eq. 1). To reach this goal, it is necessary to compute the relative weights of the different contributions. For each bootstrap iteration, the list of relevant features extracted as described in the previous section are used to compute the number of features that are associated with mildness or severity for the different frequency ranges. For instance, $n$ corresponds to the number of features associated with $r$ the mild phenotype coming from Boolean features computed for variants in the frequency range [0.1%, 1%]. A feature is considered associated with the mild phenotype when its coefficient in the LASSO model estimated in step 1 is negative, i.e. it contributes to the prediction of the phenotype adjusted by age and sex in the direction of a phenotype less severe than what expected at that particular age and sex. The same rule, applied to the corresponding Boolean representation, is used to define the other feature-counts appearing in Eq. (1). The weighting factors in Eq. (1) were estimated as the ones that maximize the Silhouette coefficient of the separation between the clusters of patients more/less severe than expected. The minimization was performed with weighting factors restricted to the following ranges: $F_{LF} \in [1, 4]$, $F_{R} \in [2, 8]$, and $F_{UR} \in [5, 100]$.

This procedure returns three optimal values for the weighing factors associated with each bootstrap iteration.

**Step 3: IPGS definition**

In this step, the data extracted at each bootstrap iteration in steps 1 and 2 are combined to define the IPGS. First, for each of the Boolean features, of all the 12 representations, the number of times this feature was selected in the 100 bootstrap iterations is computed. Then, the entire bootstrap procedure is repeated using random input phenotypes, and the 5th percentile of the number of times that a feature is associated with a random phenotype is estimated. This threshold, computed separately for each Boolean representation, was used to select which Boolean features are included in the final model. As no significant association is expected among the Boolean features and the random phenotype, the threshold of the 5th percentile is expected to exclude with a 95% level of confidence the possible false positive associations. For the GEN-COVID cohort, the features selected correspond to ~ 4.4% of the initial number of Boolean features. The weighing factors in Eq. (1) were computed as the median values of the estimates obtained in the 100 bootstrap iterations.

**Step 4: Training of the predictive model based on age, sex, and IPGS**

The procedure described in the previous sections completely defines how to calculate the IPGS. The predictive model of the binary COVID-19 severity (hospitalized patients with any form of respiratory support versus all other patients) was defined as a logistic model that uses as input features IPGS, age, and sex. It should be noted that in steps 1-3, only patients that deviates from their expected severity based on age and sex were used. The procedure was designed to isolate the genetic basis of COVID-19 severity. Instead, in this final step, IPGS, age and sex are combined to predict the actual COVID-19 severity, i.e. hospitalized patients with any form of respiratory support. To prevent overfitting, the model was fitted using 466 samples different from the training set adopted in steps 1–3. 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. The percentile normalization of the IPGS scores is performed within each cohort. An alternative logistic model that used as input features only age and sex was also fitted on the same training set. The comparison between the two models is intended to evaluate if the genetic information summarized in the IPGS improves the prediction of severity compared to a model based on age and sex alone. A further logistic regression model is fitted by only considering the IPGS variable.

**Model testing**

The training procedure returned 2 logistic models to be compared: one using as input features only age and sex, and the other one using as input features age, sex, and the IPGS. These models were tested, without any further adjustment, using other cohorts of European ancestry. The performances of the two models, with and without IPGS, were evaluated and compared in terms of accuracy, precision, sensitivity,
and specificity. The increases of the performances are evaluated with respect to the performances of a model where the values of the IPGS feature have been shuffled, by computing the p value on the empirical null distribution. In addition, the empirical probability density function of IPGS has been estimated for the severe and non-severe patients of the cohort including both train and test sets and a t test is carried out to evaluate whether the means of the two distributions were significantly different. As a further evaluation of the importance of the IPGS score on the severity prediction, univariate logistic regression models using as independent variables age (continuous represented in decades), sex, and IPGS were fitted to the dataset that combines both the training set and the testing sets for a total of 2240 patients. These models were used to estimate the odds ratios and the p values of the association with the severe phenotype. Furthermore, a multivariable logistic regression was fitted using IPGS, age, and sex together. Finally, a multivariable logistic regression was performed using as predictor variables: IPGS, age, sex and comorbidities (congestive/ischemic heart failure; asthma/COPD(OSAS; diabetes; hypertension; cancer). This latter model has been fitted in the training set, where the information on comorbidities was available.

**Pathway analysis**

Pathway analysis was made using a ranked GSEA approach (Subramanian et al. 2005; Mootha et al. 2003), modified according to the specificity of our data. The metrics for gene ranking was calculated on the basis of the results of feature selection models, weighting in each Boolean feature both beta values and the number of bootstrap iterations where it was found significantly associated with severity/mildness (Supplementary Tables 3–6). All the Boolean features that were found significant in at least one of the models were included in the list. As the sign of beta depends on which allele is taken as reference (which is relative for common variants), we decided to use absolute beta values for all the features. To also weight the importance according to variant frequency, we used the $F$ values from the IPGS score for the four categories (ultra-rare 5, rare 4, low frequency 2, common 1). Finally, we summed all the weights of each Boolean feature by gene. Briefly,

$$W_{\text{geneA}} = \sum_{\text{Features Gene A}} \text{ABS}(\text{mean}\beta) \times \text{count} \times F.$$  

Pathway enrichment analysis was made using the GSEA-preranked module (v. 7.2.4) of the GenePattern platform (Wang et al. 2010), on several pathway categories (BIOCARTA, KEGG, REACTOME, GOBP, HALLMARKS, C7 and C8), limiting the size of gene sets to the 10-300 range and performing 10,000 permutations. The networks showing similarity of significant pathways were built using the EnrichmentMap algorithm (Subramanian et al. 2005) in the Cytoscape suite (v. 3.8.2) (Merico et al. 2011; Shannon et al. 2003). Parameters used for network creations are: Jaccard Overlap Combined Index ($k$ constant = 0.5), edge cutoff 0.05.

**Website and data distribution**

The coordination of international partners has been possible through the Host Genetics Initiative (HGI) (https://www.covid19hg.org/projects/).

Results can be shared through the Gen-Covid website (https://sites.google.com/dbm.unisi.it/gen-covid).

**Code availability**

Data analyses were performed using Python with the Scipy ecosystem (Virtanen et al. 2020), and the scikit-learn library (Pedregosa et al. 2011). Statistical association was done with the statsmodel Python library. The code is freely available at the github repository: https://github.com/gen-covid/pmm.

**Results**

**The post-Mendelian paradigm for COVID-19 modelization for combining interpretability with predictivity based on ultra-rare, rare, low-frequency, and common variants**

The aim of the present study was to develop an easily interpretable model that could be used to predict the severity of COVID-19 from host genetic data. Patients were considered severe when hospitalized and receiving any form of respiratory support. The focus on this target variable is motivated by the practical importance of rapidly identifying which patients are more likely to require oxygen support, in an effort to prevent further complications. Interpretability has been a guiding principle in the definition of the machine-learning model, as only a readily interpretable model can provide useful and reliable information for clinical practice while also contributing significantly to diagnostic, and therapeutic targeting. The high dimensionality of host genetic data poses a serious challenge to evident and reliable interpretability. So far, the development of a robust predictive model able to make a direct association between single variants and disease severity grading based on an accurate analysis of the vast number of host genetic variants compared to a much smaller number of individual patients has proven to be too complex and ultimately unreliable. To address the complexity with predictive reliability, an enriched gene-level representation of host genetic data was modeled in a
machine-learning framework. The complexity of COVID-19 immediately suggests that both common and rare variants are expected to contribute to the likelihood of developing a severe form of the disease. However, the contribution of common and rare variants to the severe phenotype is not expected to be the same. A single rare variant that impairs the protein function might cause a severe phenotype by itself after viral infection, while this is not so probable for a common polymorphism, which is likely to have a less marked effect on protein functionality. These observations led to the definition of a score, named IPGS, that includes data regarding the variants at different frequencies:

\[
\text{IPGS} = (n_C^c - n_C^m) + F_{LF} \cdot (n_{LF}^s - n_{LF}^m) + R \cdot (n_R^s - n_R^m) + F_{UR} \cdot (n_{UR}^s - n_{UR}^m).
\]

In Eq. (1), \( n \) variables are used to indicate the number of input features of the predictive model that promote the severe outcome (superscript s) or that protect from a severe outcome (superscript m) and with genetic variants having Minor Allele Frequency (MAF) \( \geq 5\% \) (common, subscript C), \( 1\% < \text{MAF} \leq 5\% \) (low-frequency, subscript LF), \( 0.1\% < \text{MAF} \leq 1\% \) (rare, subscript R), and MAF < 0.1\% (ultra-rare, subscript UR). The features promoting or preventing severity were identified by an ensemble of logistic models, as described in the next section. The weighting factors \( F_{LF}, \) \( R, \) and \( F_{UR} \) were included to model the different penetrant effects of low-frequency, rare, and ultra-rare variants, compared to common variants. Thus, the 4 terms of Eq. (1) can be interpreted as the contributions of common, low-frequency, rare, and ultra-rare variants to a score that represents the genetic propensity of a patient to develop a severe form of COVID-19.

**Feature selection and gene discovery**

The definition of the single terms of the IPGS formula requires 4 separate steps (Fig. 1): (1) the definition of a severity phenotype adjusted by age and sex; (2) the conversion of genetic variants into Boolean features that represent the presence of variants in different frequency ranges in each gene; (3) the selection of those features that are associated with disease severity; and (4) the optimization of the weighting factors appearing in Equation 1. These 4 steps were executed using data from a training set extracted from the GEN-COVID dataset (90% of the patients, \( n=1780 \), see Methods). The phenotype adjusted by age and sex was computed using an ordered logistic regression model, with the purpose of facilitating the extraction of features associated with the genetic basis of COVID-19 severity (Fig. 1B). The conversion of genetic variants into Boolean features led to the definition of 12 separate sets of input features (Supplementary Table 2). The set of input features “ultra-rare_autosomal dominant” (UR_AD) is designed to represent in a binary way an autosomal dominant hereditary model associated with variants with MAF lower than 0.1\%, i.e. these Boolean features are equal to 1 for genes presenting at least one variant in this frequency range. Similarly, the set of input features “ultra-rare_autosomal recessive” (UR_AR) and “ultra-rare_X-linked” (UR_X) were designed to describe the autosomal recessive and X-linked models of inheritance of ultra-rare variants. Analogous principles were used for rare and low-frequency variants. In the case of common variants, the same 3 sets of Boolean features representing the autosomal dominant, autosomal recessive, and X-linked models of inheritance were used. However, instead of simply defining the binary variables as “absence/presence of variants”, the absence/presence of variant combinations was tested (Fig. 1C).

The Boolean representation of the genetic variability described in the previous paragraph significantly reduces the dimensionality of the problem. However, the number of input features is still orders of magnitude higher than the number of patients that can be reasonably collected for training the model. To reduce the number of input features, a feature selection strategy based on logistic models with the validated Least Absolute Shrinkage and Selection Operator (LASSO) regularization was employed (Fig. 1D). The aim of LASSO regularization is to minimize (shrink) the number of coefficients of the model, consequently minimizing the number of input features used for predicting outcomes. Separate logistic models with LASSO regularization were trained for the 12 sets of Boolean features for predicting COVID-19 severity, allowing us to identify the relevant features for each set. About 4% of the cumulative tested features were found to contribute to COVID-19 variability in severity (Fig. 1D). Twenty-six percent identity match between extracted genes and known viral susceptibility genes was found (Supplementary Table 12). We further investigate extracted genes using the Human Gene Connectome (HGC). Interestingly, of the 4943 genes of our model that are mapped in HGC, 4401 (89%) are biologically significantly connected (\( p < 0.05 \)) and 2847 (57.6%) with a degree of 0 (overlap) or 1 (direct connection) with one of the genes of the three Core Lists (Supplementary Table 13).

**Biological interpretability of extracted genetic features**

Selected genes contributed by ultra-rare, rare, low-frequency variants, or common variants (Fig. 2A–D and Supplementary Tables 3a–g). Specifically, 54% contributed by only one, 29% by two, 11% by three, and 6% by four types of variants. Around 25% of the genes were sex-specific. The latter group includes either genes located on the X chromosome, such as TLR7 and TLR8 in males, or genes regulated...
in opposite directions by androgens and estrogens when contributing with less penetrant common variants, such as p.L412F in TLR3 and p.D603N in SELP gene (Fig. 2A–D).

Among the extracted ultra-rare variants there was a group of genes, such as TLR3, TLR7 and TICAM1, already shown to be directly involved in the Mendelian-like forms of COVID-19 (Fig. 2A and Supplementary Table 3a, b). Furthermore, another group of genes are natural candidates because of their function: these include the ACE2 shedding protein ADAM17, CFTR-related genes, genes involved in glycolipid metabolism, genes expressed by cells of the innate immune system, and genes involved in the coagulation pathway. Finally, a group of genes led by ACE2 (if affected by ultra-rare variants) confers protection from the severe disease. This group includes several genes whose mutations are responsible for auto-inflammatory disorders.

Among the rare variants extracted, we identified some genes as candidates for COVID-19 severity, including TLR5 and SLC26A9 as well as other genes involved in the inflammatory response (Fig. 2B and Supplementary Table 4a, b).

Among the low-frequency variants extracted, we identified some genes associated with either severity or protection from severe COVID-19 that are linked to the CFTR pathway (e.g., PSMA6) as well as specific genes involved in the immune response (e.g., NOD2) (Fig. 2C and Supplementary Table 5a, b).

The model was also able to identify a group of extracted common variants already shown to be linked to either severe
or mild COVID-19 (Fig. 2D and Supplementary Table 6a, b). Among them are the L412F TLR3 and D603N SELP polymorphisms, already reported to be associated with the severe disease (Croci et al. 2021; Fallerini et al. 2021b) and several coding polymorphisms in Linkage Disequilibrium (LD) with already reported genomic SNP, such as the ABO blood group, OAS1-3 genes, PPP1R15A gene and others (Elhabyan et al. 2020). In conclusion, considering their functions, genes involved in the immune and inflammatory responses, or those involved in the coagulation pathway and NK and T cell receptor, are to be considered natural candidates for severe or mild COVID-19.

**Integrated PolyGenic Score definition**

The Boolean features selected by the LASSO logistic models were used to calculate ten variables in Eq. (1) (Fig. 3A, B). The corresponding weights (F variables) were defined by optimizing the separation between severe and mild cases as offered by the IPGS formula. The optimization was measured using the Silhouette coefficient, and the optimal values were computed using a grid-search approach over a predefined grid (\(F_{L} \in [1, 4], F_{R} \in [2, 8], \) and \(F_{UR} \in [5, 100]\)).

This optimization returned values of 2, 4, and 5 for the low-frequency, rare, and ultra-rare variants, respectively (Fig. 3C, D).

**Pathway analysis**

To understand the biological mechanisms underlying the variability of disease, we performed a pathways analysis of the genes carrying variants discovered in the feature selection described above. The features obtained with this approach do not have the same predicted impact and are not discovered with the same confidence. Therefore, we decided to perform a rank-based pathway analysis, with genes with the highest impact and confidence ranking highest in the list, rather than a simple over-representation approach. We ranked all the genes that were found to be significantly associated with severity/mildness in at least one bootstrap repetition, based on a score that takes into account three parameters: average coefficient in the LASSO models selecting the feature, number of significant bootstrap results, and the \(F\) correction factor for the frequency category used in the IPGS (detailed in the Methods section below). For genes with more than one significant Boolean feature, we summed up the scores of each feature. Gene Set Enrichment Analysis (GSEA) was then performed using two separate ranked gene lists (Supplementary Table 7) for females and males, followed by the generation of similarity networks using EnrichmentMap (Fig. 4A). The usage of rank-based search method allows to identify statistically significant pathways starting from extensive list genes, as each gene is associated with its specific importance. Although no pathways satisfied the 0.25% FDR threshold normally required for standard GSEA analyses, the set of pathways considered significant using more relaxed thresholds on \(p\) values were shown to group in meaningful modules, providing useful information on pathogenetic mechanisms and on the genes that could explain how they can be affected. The network of all the pathways significantly enriched in both females and males ranked gene lists (\(p < 0.01, n = 25\)) is depicted in Fig. 4B, while the network of all the pathways enriched in either females or males with a more stringent \(p\) value (\(p < 0.005, n = 100\)) is shown in Fig. 4C. Detailed information on the names of the pathways and \(p\) values of enrichment is reported in Supplementary Figures 1 and 3. For the most representative pathways of each network, the heatmaps of the genes with their weights of association to disease variability are shown in Fig. 4D and Supplementary Figures 2 and 4, while gene lists and gene weights for all the significant pathways are reported in Supplementary Table 8.

**COVID-19 post-Mendelian model predictivity**

The functional interpretation of the variants identified by the feature selection approach, complemented by the strong link between the involved human biological pathways and COVID-19 pathogenicity, support the hypothesis that the IPGS score developed here may contribute significantly to predicting the severity of COVID-19. This hypothesis was tested using a logistic regression model that predicts COVID-19 severity based on age, sex and the IPGS (after percentile normalization). The training set is composed of 466 patients not included in the training set previously exploited for the IPGS feature engineering. The model’s performance was then tested using three independent cohort sets of European ancestry (Fig. 5A). The model exhibited an overall accuracy of 0.73, precision equal to 0.78, with a sensitivity and specificity of 0.72 and 0.75, respectively. Noteworthy, all the aforementioned metrics are higher than the corresponding values obtained using a logistic model that adopted as input features only age and sex. The increase in performances of the model with IPGS suggests that this score indeed confers significant additional (genetic) information for predicting COVID-19 severity compared to only age and sex. The increase of the performances is statistically significant (\(p\) value < 0.05 for accuracy, precision, sensitivity, specificity) with respect to the distribution of performances for an ensemble of models where the IPGS feature has been randomized (Fig. 5C lower left). A third logistic regression model fitted with IPGS alone, shows performances well above the random guess. Furthermore, the empirical probability density function of IPGS scores (Fig. 5C right) has been estimated for the severe and non-severe patients of the cohort.
including both training and testing sets. It is worth noting
the shift on the right of the IPGS distribution for the severe
patients, with significant p value (< 0.001) for the t test of
mean difference. This difference between severe and non-
severe cases is preserved for the male and female cohorts
when analyzed separately (p values < 0.001 and 0.024,
respectively).

In line with the results obtained using the overall test set,
the model including IPGS, age, and sex performed better than
the model considering only sex and age as inputs, in each
of the testing cohorts, separately (Fig. 5D). The increase in
performance was systematically observed throughout all the
cohorts: on average + 1.33% for accuracy, + 1% for precision,
+ 1.33% for sensitivity, + 1.67% for specificity. Considering
the difference in phenotype classification inherent to a com-
parison among various international cohorts, and the genetic
variability among different European sub-populations, the
consistent increase in performances observed for the model
with IPGS demonstrates that this score provides a robust
index for predicting COVID-19 severity. As a further test for
the importance of the IPGS score for predicting COVID-19
severity, the univariate logistic models were used on the over-
all set including both train and test cohorts to estimate the
OR of severe COVID-19 for IPGS, age, and sex, separately.
The test confirmed that severity was associated with IPGS,
showing an OR of 2.32 (p < 0.001, 95% confidence interval
[1.79, 3.01]) with age, measured in decades, and sex, having
OR of 1.89 (p < 0.001, 95% confidence interval [1.79, 2.00])
and 2.99 (p < 0.001, 95% confidence interval [2.58, 3.46])
respectively (Fig. 5E). The multivariate logistic regression
using sex, age, and IPGS together, provided similar results
reported in Supplementary Table 9 confirming the goodness
of the regressors’ OR. When adjusting for comorbidities, in
the train cohort where the comorbidities were available, with a
multivariable logistic model, OR of IPGS was 2.46 (p = 0.05,
95% confidence interval [1.15, 5.25]) as shown in Supple-
mentary Table 10. This result further confirms that IPGS is a
reliable predictor of COVID-19 clinical severity.
Advantages of IPGS and clinical interpretability of connected features

We then wanted to compare the clinical outcome with the probability of severity obtained from three different models: IPGS alone, sex-age alone or combined model (represented as heatmap in Fig. 6). It appears evident that in a subset of patients, the 2 models based on sex-age alone and IPGS alone have a discordant prediction (left and right end of dendrogram in Fig. 6A). In these cases, IPGS appears to be a relevant predictor of severity (Fig. 6A). This is in accordance with the above-presented logistic regression analysis (Fig. 5E) that shows IPGS having an OR of 2.32 for severity. Moreover, the list of features on which the IPGS score is built, represent a biological handle for pathophysiological mechanisms and possible personalized adjuvant treatments.

For example, three male patients, within two distinct age ranges (46–50, 51–55) (panels B, C and D) with severe outcome (intubation and CPAP) are imperfectly represented by the sex-age model (probability of severity from 0.52 to 0.66) and better represented by the IPGS model (probability of severity from 0.91 to 0.95). The detected genetic variants that would allow to clinically consider putative personalized treatments in similar cases are: (1) TLR7 ultra-rare mutation indicating to consider possible adjuvant treatment with IFN gamma administration (Fallerini et al. 2021a); (2) homozygosity 603Asn in SELP gene suggesting putative adjuvant treatment with anti-selectin P autoantibodies (e.g. Crizanlimzubam) (Fallerini et al. 2021b) and (3) polyQ longer than 23 in AR gene suggesting to consider possible adjuvant treatment with testosterone (Baldassarri et al. 2021a).

In a female patient, within age range 31–35, the sex-age model showed a probability of severity of 0.17 (panel D) while the IPGS score was 336 corresponding to a probability of severity from 0.91 to 0.95). The detected genetic variants that would allow to clinically consider putative personalized treatments in similar cases are: (1) TLR7 ultra-rare mutation indicating to consider possible adjuvant treatment with IFN gamma administration (Fallerini et al. 2021a); (2) homozygosity 603Asn in SELP gene suggesting putative adjuvant treatment with anti-selectin P autoantibodies (e.g. Crizanlimzubam) (Fallerini et al. 2021b) and (3) polyQ longer than 23 in AR gene suggesting to consider possible adjuvant treatment with testosterone (Baldassarri et al. 2021a).

For example, three male patients, within two distinct age ranges (46–50, 51–55) (panels B, C and D) with severe outcome (intubation and CPAP) are imperfectly represented by the sex-age model (probability of severity from 0.52 to 0.66) and better represented by the IPGS model (probability of severity from 0.91 to 0.95). The detected genetic variants that would allow to clinically consider putative personalized treatments in similar cases are: (1) TLR7 ultra-rare mutation indicating to consider possible adjuvant treatment with IFN gamma administration (Fallerini et al. 2021a); (2) homozygosity 603Asn in SELP gene suggesting putative adjuvant treatment with anti-selectin P autoantibodies (e.g. Crizanlimzubam) (Fallerini et al. 2021b) and (3) polyQ longer than 23 in AR gene suggesting to consider possible adjuvant treatment with testosterone (Baldassarri et al. 2021a).

In a female patient, within age range 31–35, the sex-age model showed a probability of severity of 0.17 (panel D) while the IPGS score was 336 corresponding to a probability of severity from 0.91 to 0.95). The detected genetic variants that would allow to clinically consider putative personalized treatments in similar cases are: (1) TLR7 ultra-rare mutation indicating to consider possible adjuvant treatment with IFN gamma administration (Fallerini et al. 2021a); (2) homozygosity 603Asn in SELP gene suggesting putative adjuvant treatment with anti-selectin P autoantibodies (e.g. Crizanlimzubam) (Fallerini et al. 2021b) and (3) polyQ longer than 23 in AR gene suggesting to consider possible adjuvant treatment with testosterone (Baldassarri et al. 2021a).
Discussion

The most robust and traditional method for feature selection in complex disorders is Genome-wide association studies (GWASs). GWAS focuses on common variants only whose effects are small. The method is based on the comparison of about 700,000 genomic SNPs, mostly non-coding, in cases/controls. The coverage of coding SNPs is usually only through imputed data by Linkage Disequilibrium. The method needs ten/hundred of thousands subjects. The missing heritability of this method is rare variants. Until now, 35 loci have been identified (Kosmicki et al. 2021) (WES/WGSHGI Working Group https://www.covid19hg.org/blog/2021-09-27-september-20-2021-meeting/). None of the above methods until now has been able to reach predictivity of COVID-19 severity useful in clinical practice.

The new proposed model takes into account both common and rare variants and has the ability to extract significant COVID-19 features (genes/variants) using a set of relatively low numbers of subjects. Furthermore, employment of oligo-asymptomatic SARS-CoV-2-infected subjects as

Fig. 3 Integrated PolyGenic Score Definition. A The model is based on the comparison of Boolean features of severity versus Boolean features of mildness. B Graphic representation of the IPGS formula used for this model. C Principle for the calibration of different weighting factors based on the separation of severe and mild cases. D The obtained value for low-frequency, rare, and ultra-rare, being $F = 1$ for common variants. Common variants are indicated as common haplotypes since they are intended as combinations of coding variants within a single gene (see Fig. 1C and the Material and methods section)
Fig. 4 Pathway enrichment analysis of the genes associated with disease severity/mildness. A Workflow of the analysis. Genes corresponding to Boolean features found to be associated at least once were ranked based on a composite score and subjected to Gene Set Enrichment Analysis. Two separate ranked gene lists for females (7317 genes, weight range 3 × 10⁻⁵-561) and males (7325 genes, weight range 7 × 10⁻⁵-452) were used. The list of significant pathways was analysed and presented as a similarity network: B Similarity network of the pathways with a significant enrichment both in females and males (p < 0.005). D Heatmaps of the genes belonging to a selection of pathways of interest. The color gradient represents the weight of each gene, calculated and described in methods. Please note high ranking of TLR genes (TLR5, TLR8, TLR3 and TLR7) in the pathway of Response to Mechanical Stimulus, CFTR gene in Recognition for Clathrin-mediated endocytosis, RNASEL, TYK2, OAS1 and OAS3 genes in Interferon alpha–beta signaling. Note also the presence of the relevant pathway of Exhaust vs Memory CD8 T cell Up that also includes TLR7 gene.
controls instead of general population is providing a significant advantage in accuracy. Extracted genes were prioritized by matching with known viral susceptibility genes and transcriptomic data (Supplementary Table 12). Identity match was found for more than 25% of genes, even if the transcriptomic list is far to be representative of whole transcript alteration in the several tissues/organs involved in COVID-19 since most data comes from 2 tissues only (lung and swab). More than 55% of genes have a direct connection (degree 0 and 1 in HGC) with viral susceptibility genes, including those from GWAS loci and the genes in eQTL with them by GTEx analysis (Supplementary Table 13). Finally, up to 89% of genes are biologically significantly connected ($p < 0.05$) with the list above, supporting the validation of extracted genes.

Fig. 5  Model predictivity. A The post-Mendelian model was trained using a sample of 466 patients from the GEN-COVID cohort n.2 and Swedish cohort (having cases only) and tested with three additional European cohorts from UK, Germany and Canada. B A logistic regression model was used for severity prediction. Severity was defined mainly on the basis of hospitalization versus not hospitalization. Hospitalized cases without respiratory support were included in controls. TN = true negative; TP = true positive; FN = false negative; FP = false positive. C When the IPGS is added to age and gender as a regressor, the performances of the model increase: accuracy +1%, precision +1%, sensitivity +2%, specificity +1%. These increases are statistically significant ($p$ value <0.05 for accuracy, precision, sensitivity and specificity) with respect to the null distribution obtained by randomizing the IPGS. The performances of the model built with IPGS alone are all above the random guess. In addition, on the right, we reported the distributions of the IPGS for severe and non-severe patients. D In the three tested cohorts, when the IPGS is added to age and sex as a regressor, all the performances increase: the accuracy up to +2%, the precision up to +1%, the sensitivity up to +3%, and the specificity up to +2%. We conclude that IPGS is able to improve prediction of clinical outcome in addition to the well-established powerful factors of age and sex. E The univariate logistic regression models fitted on the cohort including both train and test, confirmed that the IPGS is associated with severity with an odds-ratio (OR) of 2.32, while age (continuous in decades) and sex have an OR of 1.89 and 2.99, respectively.
The importance of combination of rare and low-frequency variants has already been demonstrated to contribute to the prediction of susceptibility in other complex disorders (Marouli et al. 2017; International Multiple Sclerosis Genetics Consortium 2018). Here, we further expand this approach while demonstrating that ultra-rare, rare, low-frequency, as well as common variants contribute to the likelihood of developing a severe form of COVID-19. Furthermore, we included in our analyses a calibration of the relative weight of the variants vis-a-vis their impact on disease severity: a single ultra-rare variant might well by itself cause a severe phenotype of COVID-19, while this is less probable for a common polymorphism, one that is likely to have a markedly less direct effect on protein functionality. We performed a first modelization of COVID-19 genetics using both rare and common variants (Picchiotti et al. 2021). Since feature selection methodologies are generally sensitive to allele frequency, the extraction was performed separately...
for rare (MAF < 1/100) and common (MAF > 1/100) variants. However, the methodology revealed the insight that low-frequency variants (MAF from 1 to 5%) were disadvantaged if selected together with common ones. Furthermore, for extracting Mendelian-like genes a threshold of MAF < 0.1% (ultra-rare variants) appeared more effective than MAF < 1% and all mutations in the TLR7 gene that proved to have loss of function had indeed MAF < 1/1000 (Fallerini et al. 2021a). The model we arrived at, now considers separately ultra-rare, rare, low-frequency, and common variants.

Similar to the classical PRS (Polygenic Risk Score), the proposed IPGS (Integrated PolyGenic Score) may prove reliable for assessing the probability of severe COVID-19 following infection by SARS-CoV-2 (Mars et al. 2020). While PRS is based on common polymorphisms found at the genomic level with the majority of loci potentially conferring risk being not easily interpretable due to the uncertainty of linked genes, IPGS allows immediate biological interpretation because it only includes coding variants. Furthermore, as opposed to PRS, IPGS relies on both polymorphisms and rare variants is capable of differentially weighting features in an indirectly proportional way in respect to frequency, and therefore, to protein impact. Each patient indeed is assigned both a number and the list of her/his common and low-frequency polymorphisms relevant to COVID-19 supported by medically actionable information and of rare and ultra-rare variants conferring either risk of severity or protection from severe disease. Drawing on the entire picture presented through IPGS analysis, personalized adjuvant therapy could be envisaged. At the time of writing, a platform trial based on genetic markers is being discussed with the Italian Medicines Agency (EudraCT Number: 2021-002817-32).

Within 25 reported genomic SNPs demonstrably related to COVID-19 susceptibility/severity, 5 were reported to be in LD with coding variants (COVID-19 Host Genetics Initiative et al. 2021; Covid19hg.org 2021). The model presented here might provide useful information for uncovering the identity of the gene/coding variants responsible for COVID-19 susceptibility/severity linked to these genomic SNPs (Fig. 2D). For example, on chromosome 12, the genes mapping to the locus tagged by rs10774671 (COVID-19 Host Genetics Initiative et al. 2021) are both OAS1 and OAS3. In OAS3, the coding variant is an Arginine to Lysine substitution (rs1859330) in high LD (0.8) with the tag SNP. This polymorphism was already associated with viral infection (Tan et al. 2017) based on the presence of Lysine having been shown to lead to a decreased INF-γ production. In OAS1, the haplotype (including 4 missense variants: G162S, A352T, R361T, and G397R), the splicing variant 1039-G>A (the reported genomic polymorphism itself), and the truncating mutation T359fs*26 are associated with severity and predicted to impair OAS1 function. Both OAS1 and OAS3 induce RNASEL, which in turn exerts antiviral activity. Further support for the role of the OAS/RNASEL axis is indicated by the presence of ultra-rare recessive variants.

This innovative approach allowed us to better select genes located on the X chromosome related to COVID-19 that affect males and females in opposite ways (Fig. 2A and Supplementary Tables 3 and 6). Interestingly, many of these genes were previously confirmed or hypothesized to escape the X chromosome inactivation. With respect to these genes, females produce twice the levels of protein in comparison with males. Mutations in hemizygous state in males and heterozygous state in females appear silent until SARS-CoV-2 infection occurs. For example, TLR7 and TLR8 are selected for ultra-rare and associated with severity in males and with protection from severe disease in heterozygous females. We know that the activation of TLR7/8 induces the production of type 1 and type 2 IFN as well as pro-inflammatory cytokines, where the production defect in hemizygous males leads to severe COVID-19. However, an excess of the sensor can also lead to damage from hyperinflammation. Therefore, the condition of carrier females is the more favorable state and has in fact been associated with mild COVID-19 (Subramanian et al. 2006).

Pathway analysis pointed to the relevance of obvious actors in COVID-19 pathology, such as immune cells and interferon signaling, but also to the important role of specific organs (brain, digestive tract, kidney, reproductive system) and functions (metabolism of lipids and steroids). The pathways identified through GSEA analyses reflected the multi-organ nature of the disease. In addition, our analyses reveal new candidate determinants of disease variability. The four pathways linked to cilium motility suggest a role for ciliated cells of the respiratory tract (and possibly others) in antiviral defense. The functionality of the clathrin-mediated endocytosis pathway may likely affect viral entry (Bayati et al. 2021). Likewise, endoplasmic reticulum associated protein degradation (ERAD), which is linked to autophagy and SARS-CoV-2 life-cycle (Reggiori et al. 2010), may also be relevant. Other pathways with a less obvious but potentially interesting role in the disease include cell adhesion and mechanical stimulus signaling.

The strong link between the involved human biological pathways and COVID-19 pathogenicity support the hypothesis that the proposed IPGS equation may contribute significantly to predicting the disease severity of COVID-19. Indeed, an overall significant increase of performance was obtained in comparison with the model based on solely on age and sex. Furthermore, the IPGS is significantly associated with severity, showing an OR of 2.46 after adjustment for age, sex, and comorbidities. This indicates that IPGS is a novel prognostic factor that should be considered in the management of COVID-19 patients.

Modelling precisely the role of the entire range of host genomics affecting disease susceptibility and severity in
COVID-19 is critical to obtaining a complete biological understanding of the aetiology and pathogenicity of COVID-19 as well as other severe complex diseases. The application of iPGS based on Machine Learning principles within a post-Mendelian model allows us to more precisely identify the gene variants at play in COVID-19 as well as their specific roles, individually and in combination. This deep dive into the genetic architecture that allows for, contributes to, or even helps prevent diseases while increasing or decreasing their impact is critical for, and directly translatable into, (personalized) medicines development as well as prevention and treatment protocols. An integrated modelling of genetic variants based on a limited patient cohort, even limited in its geographical spread, may be sufficient for the development of diagnosis, and therapies across a wider range of populations. The advantage of this iPGS post-Mendelian model is that it learns and continues to learn as well as being a model from which we can obtain insights on the fundamental architecture of human genomics when confronted with severe and complex diseases.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00439-021-02397-7.

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 D-Connect. We thank the CINECA consortium for providing computing resources and the Network for Italian Genomics (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/). We thank Uppsala Intensive Care COVID-19 research group; Tomas Luther, Anders Larsson, Katja Hanslin, Anna Gradin, Sarah Galien, Sara Bulow Anderberg, Jacob Rosén and Sten Rubertsson. We thank the patients and their loved ones who volunteered to contribute to this study at one of the most difficult times in their lives, and the research staff in every intensive care unit who recruited patients at personal risk during the most extreme conditions we have ever witnessed in UK hospitals. Whole-genome sequencing was done by Illumina in partnership with the University of Edinburgh and Genomics England and was funded by UK Department of Health and Social Care, UKRI and LifeArc. The views expressed are those of the authors and not necessarily of the DHSC, DfID, NIHR, MRC, Wellcome Trust or PHE. The Health Research Board of Ireland (Clinical Trial Network Award 2014-12) funds collection of samples in Ireland. This study owes a great deal to the National Institute of Healthcare Research Clinical Research Network (NIHR CRN) and the Chief Scientist Office (Scotland), who facilitate recruitment into research studies in NHS hospitals, and to the global ISARIC and INFACT consortia. We also acknowledge funding provided by the German Research Foundation (DFG) through the NGS competence centres (INST 37/1049-1, INST 216/981-1, INST 257/605-1, INST 269/768-1 and INST 217/988-1) and scientific support through the German COVID multomics initiative (DeCoI). Recruitment of the COMRI Study/ of part of the DeCoI samples was funded by the Klinikum rechts der Isar, Technical University Munich, Munich, Germany. We received additional support by the DFG (LU 1944/3-1, to K.U.L. and SCHU 2419/2-1, to E.C.S.), the BONFOR program of the Medical Faculty of the University of Bonn (2020-1A-15, to A.S.) and the Munich Clinician Scientist Program (MCSP, to E.C.S.).

GEN-COVID Multicenter Study (https://sites.google.com/dbm.unisi.it/gen-covid) Francesca Montagnani1-27, Mario Tumbarello1-27, Ilaria Rancan1-27, Massimiliano Fabbianni1-27, Barbara Rossetti1-27, Laura Bergantini1-27, Miriana D’Alessandro1-27, Paolo Cameli1-27, David Bennett28, Federico Anderda28, Simona Marcantonio28, Sabino Sciolletta28, Federico Franchi28, Maria Antonetta Mazzei28, Susanna Guerri28, Edoardo Conticini28, Luca Cantarini28, Bruno Frediani28, Danilo Taccioni28, Chiara Spretelli Raffaeli28, Marco Feri28, Alice Donatti28, Raffaele Scala34, Luca Guidelli34, Genni Spargi34, Marta Corridi35, Cesira Nencioni36, Leonardo Croci36, Gian Pietro Caldarelli37, Maurizio Spagnesi38, Davide Romani38, Paolo Picentini38, Maria Bandini38, Elena Desanctis38, Silvia Cappelli38, Anna Canaccini39, Agnese Verzuri39, Valentina Anemoli39, Manola Pisani39, Agostino Ognibene40, Alessandro Pancrazzi40, Maria Lorubbio40, Massimo Vaghi40, Antonella D’Arminio Monforte40, Federica Gaia Miraglia40, Mario U. Mondelli41, Massimo Girardis41, Sophie Venturelli41, Stefano Busani41, Andrea Costarizza41, Andrea Antinori41, Alessandra Vergori41, Arianna Emilizzoli41, Stefano Rusconi42-49, Matteo Siano42-49, Arianna Gabrielli42-49, Agostino Riva42-49, Daniela Francisci43, Elisabetta Schiaroli43, Francesco Piacosì43, Andrea Tommasi43, Pier Giorgio Scottoni43, Francesca Andretta43, Sandro Panese43-52, Stefano Barattì43, Renzo Scaggiante43-52, Francesca Gatti44, Saverio Giuseppe Parisi44, Francesco Castelli53, Eugenia Quiros-Roldan53, Melania Degli Antoni53, Isabella Zanella53-57, Matteo Della Monica58, Carmelo Piscopo58, Marco Capasso58-60, Roberta Russo58-60, Immacolata Andolfi58-60, Achille Iolascon58-60, Giuseppe Fiorentino61, Massimo Carella61, Marco Castori61, Filippo Acculla61, Pamela Ragg61, Rita Perna61, Matteo Bassetti62-66, Antonio Di Biagio62-66, Maurizio Sanguinetti62-66, Luca Masuccio62-66, Alessandra Guarnaccia63, Serafina Valente63, Oreste De Vivo63, Gabriella Doddatto63, Rossella Tina63, Annarita Giliberti63-65, Maria Antonietta Mencarelli63, Caterina Lo Rizzo63, Anna Maria Pinto63, Valentina Perticari63-65, Francesca Ariani63-65, Miriam Lucia Carriero63-65, Laura Di Sarno63-65, Diana Alaverdian63-65, Elena Bargagli63, Marco Mandalà63, Alessia Giorri63, Lorenzo Salerni63, Patrizia Zucchi63, Pierpao Parravicini63, Elisabetta Menatti63, Tullio Trotta64, Ferdinando Giannattasio64, Gabriella Coiro64, Fabio Lena64, Leonardo Gianluca Lacerenza64, Domenico A. Coviello64, Cristina Muggino64, Enrico Martinelli64,57, Stefano Mancarella64-67, Luisa Tavecchia64-67, Mary Ann Belli64-67, Lia Crotti64-67,81,82,83,84, Gianfranco Parat64-81, Maurizio Sanarico82, Francesco Raimondi82, Filippo Biscarini82, Alessandra Stella82, Marco Rizz82, Franco Maggiolo82, Diego Ripamonti82, Claudia Suardi82, Tiziana Bachetti83, Maria Teresa La Rovere83, Simona Sarzi-Braga83, Maurizio Bussotti83, Katia Capitanii83, Simona Dei83, Sabrina Ravaglia83, Rosangela Artuso83, Elena Andreucci83, Giulia Gori83, Angelica Pagliazzi83, Erika Fiorentino83, Antonio Perrella83, Francesca Bianchi84, Paolo Bergomi84, Emanuele Catena84, Riccardo Colombo84, Sauro Luchi84,100, Giovanni Morelli84, Paolo Petrelli84, Sarah Iacopini84, Sara Modica84, Silvia Baroni85, Francesco Vladimiro Segala85, Francesco Menichetti85, Marco Falcone85, Giusey Tiseo85, Chiara Barbier85, Tommaso Matucci85, Davide Grassi85, Franco Ferri85, Franco Marinaro85, Franco Brancati85, Antonella Vincenti85, Valentina Borgo85, Lombardi Stefania85, Marco Lenzii85, Massimo Antonio Di Pietro86, Francesca Vichi86, Benedetta Romanini86, Letizia Attala86, Cecilia Coviati86, Andrea Gabbotti86, Mené Roberto86, Umberto Zuccon86, Lucia Vietti86, Stefano Ceri86, Pietro Pinioni86, Patrizia Casprini86, Giuseppe Merla87,113,114, Gabriella Maria Squero113, Marcello Maffezzoni113,114, Raffaele Bruno116,117, Marco Vecchia116, Marta Colaneri116, Serena Ludovisi118.

WES/WGS working group within the HGI (https://www.covid19hg.org/projects/) the genetic predisposition to severe COVID-19 (SweCovid):
Guys and St Thomas’ Hospital, London, UK. (150) James Cook University Hospital, Middlesbrough, UK. (151) The Royal Liverpool University Hospital, Liverpool, UK. (152) King’s College Hospital, London, UK. (153) Royal Infirmary of Edinburgh, Edinburgh, UK. (154) John Radcliffe Hospital, Oxford, UK. (155) Addenbrooke’s Hospital, Cambridge, UK. (156) Morriston Hospital, Swansea, UK. (157) Ashford and St Peter’s Hospital, Surrey, UK. (158) Royal Stoke University Hospital, Staffordshire, UK. (159) Queen Elizabeth Hospital, Birmingham, UK. (160) Glasgow Royal Infirmary, Glasgow, UK. (161) Kingston Hospital, Surrey, UK. (162) The Tamridge Wells Hospital and Maidstone Hospital, Kent, UK. (163) North Middlesex University Hospital NHS trust, London, UK. (164) Bradford Royal Infirmary, Bradford, UK. (165) Blackpool Victoria Hospital, Blackpool, UK. (171) Manchester Royal Infirmary, Manchester, UK. (168) St George’s Hospital, London, UK. (169) Good Hope Hospital, Birmingham, UK. (170) Stepping Hill Hospital, Stockport, UK. (171) Manchester Royal Infirmary, Manchester, UK. (172) Royal Alexandra Hospital, Paisley, UK. (173) Queen Elizabeth University Hospital, Glasgow, UK. (174) Queen Alexandria Hospital, Portsmouth, UK. (175) BHRUT (Barking Havering)—Queens Hospital and King George Hospital, Essex, UK. (176) University College Hospital, London, UK. (177) Royal Victoria Infirmary, Newcastle Upon Tyne, UK. (178) Western Sussex Hospitals, West Sussex, UK. (179) Salford Royal Hospital, Manchester, UK. (180) The Royal Old-Aged Hospital, London, UK. (181) Royal Brompton Hospital, London, UK. (182) Basildon Hospital, Basildon, UK. (183) University Hospital of Wales, Cardiff, UK. (184) Broomfield Hospital, Chelmsford, UK. (185) Royal Cornwall Hospital, Truro, UK. (186) Nottingham University Hospital, Nottingham, UK. (187) Royal Hal- lamsbury Hospital and Northern General Hospital, Sheffield, UK. (188) Royal Hampshire County Hospital, Hampshire, UK. (189) Queens Hospital Burton, Burton-On-Trent, UK. (190) New Cross Hospital, Wolverhampton, UK. (191) Heartlands Hospital, Birmingham, UK. (192) Walsall Manor Hospital, Walsall, UK. (193) Stoke Mandeville Hospital, Buckinghamshire, UK. (194) Sandwell General Hospital, Birmingham, UK. (195) Royal Berkshire NHS Foundation Trust, Berks- hire, UK. (196) Charing Cross Hospital, St Mary’s Hospital and Hammersmith Hospital, London, UK. (197) Dumfries and Galloway Royal Infirmary, Dumfries, UK. (198) Bristol Royal Infirmary, Bristol, UK. (199) Royal Sussex County Hospital, Brighton, UK. (200) Whiston Hospital, Prescot, UK. (201) Royal Glamorgan Hospital, Cardiff, UK. (202) King’s Mill Hospital, Nottingham, UK. (203) Fairfield General Hospital, Bury, UK. (204) Western General Hospital, Edinburgh, UK. (205) Northwick Park Hospital, London, UK. (206) Royal Preston Hospital, Preston, UK. (207) Royal Derby Hospital, Derby, UK. (208) Sunderland Royal Hospital, Sunderland, UK. (209) Royal Surrey County Hospital, Guildford, UK. (210) Derriford Hospital, Plymouth, UK. (211) Croydon University Hospital, Croydon, UK. (212) Victoria Hospital, Kirkcaldy, UK. (213) Milton Keynes University Hospital, Milton Keynes, UK. (214) Barnsley Hospital, Barnsley, UK. (215) York Hospital, York, UK. (216) University Hospital of North Tees, Stockton on Tees, UK. (217) University Hospital Wishaw, Wishaw, UK. (218) Whittington Hospital, London, UK. (219) Southend Hospital, Bristol, UK. (220) The Royal Papworth Hospital, Cambridge, UK. (221) Royal Gwent Hospital, Newport, UK. (222) Norfolk and Norwich University Hospital (NUHU), Norwich, UK. (223) Great Ormond St Hospital and UCL Great Ormond St Institute of Child Health NIHR Biomedical Research Centre, London, UK. (224) Airedale General Hospital, Keighley, UK. (225) Aberdeen Royal Infirmary, Aberdeen, UK. (226) Southampton General Hospital, Southampton, UK. (227) Russell’s Hall Hospital, Dudley, UK. (228) Rotherham General Hospital, Rotherham, UK. (229) North Manchester General Hospital, Manchester, UK. (230) Basingstoke and North Hampshire Hospital, Basingstoke, UK. (231) Royal Free Hospital, London, UK. (232) Hull Royal Infirmary, Hull, UK. (233) Hartlepool Hospital, London, UK. (234) Chesterfield Royal Hospital Foundation Trust, Chesterfield, UK. (235) Barnet Hospital, London, UK. (236) Aintree University Hospital, Liverpool, UK. (237) St James’s University Hospital and Leeds General Infirmary, Leeds, UK. (238) Glen Clwyd Hospital, Bodelwyddan, UK. (239) University Hospital Crosshouse, Kilmarnock, UK. (240) Royal Bolton Hospital, Bolton, UK. (241) Princess of Wales Hospital, Llantrisant, UK. (242) Pilgrim Hospital, Lincoln, UK. (243) Northumbria Healthcare NHS Foundation Trust, North Shields, UK. (244) Ninewells Hospital, Dundee, UK. (245) Lister Hospital, Stevenage, UK. (246) Bedford Hospital, Bedford, UK. (247) Royal United Hospital, Bath, UK. (248) Royal Bournemouth Hospital, Bournemouth, UK. (249) The Great Western Hospital, Swindon, UK. (250) Watford General Hospital, Watford, UK. (251) University Hospital North Durham, Darlington, UK. (252) Tameside General Hospital, Ashton Under Lyne, UK. (253) Princess Royal Hospital Shrewsbury and Royal Shrewsbury Hospital, Shrewsbury, UK. (254) Arrowe Park Hospital, Wirral, UK. (255) The Queen Elizabeth Hospital, King’s Lynn, UK. (256) Royal Blackburn Teaching Hospital, Blackburn, UK. (257) Poole Hospital, Poole, UK. (258) Medway Maritime Hospital, Gillingham, UK. (259) Warwick Hospital, Warwick, UK. (260) The Royal Marsden Hospital, London, UK. (261) The Princess Alexandra Hospital, Harlow, UK. (262) Musgrove Park Hospital, Taunton, UK. (263) George Eliot Hospital NHS Trust, Nuneaton, UK. (264) East Surrey Hospital, Redhill, UK. (265) West Middlesex University Hospital, Isleworth, UK. (266) Warrington General Hospital, Warrington, UK. (267) Southport and Formby District General Hospital, Ormskirk, UK. (268) Royal Devon and Exeter Hospital, Exeter, UK. (269) Macclesfield District General Hospital, Macclesfield, UK. (270) Borders General Hospital, Melrose, UK. (271) Birmingham Children’s Hospital, Birmingham, UK. (272) William Harvey Hospital, Ashford, UK. (273) Royal Lancaster Infirmary, Lancaster, UK. (274) Queen Elizabeth the Queen Mother Hospital, Margate, UK. (275) Liverpool Heart and Chest Hospital, Liverpool, UK. (276) Darlington Memorial Hospital, Darlington, UK. (277) Southend University Hospital, Westcliff-on-Sea, UK. (278) Raigmore Hospital, Inverness, UK. (279) Salisbury District Hospital, Salisbury, UK. (280) Peterborough City Hospital, Peterborough, UK. (281) Ipswich Hospital, Ipswich, UK. (282) Hereford County Hospital, Worcester, UK. (283) Furness General Hospital, Barrow-in-Furness, UK. (284) Forth Valley Royal Hospital, Falkirk, UK. (285) Torbay Hospital, Torquay, UK. (286) St Mary’s Hospital, Newport, UK. (287) Royal Manchester Children’s Hospital, Manchester, UK. (288) Royal Cornwall Hospital, Truro, UK. (289) Queen Elizabeth Hospital Gateshead, Gateshead, UK. (290) Kent & Canterbury Hospital, Canterbury, UK. (291) James Paget University Hospital NHS Trust, Great Yarmouth, UK. (292) Darenth Valley Hospital, Dartford, UK. (293) The Alexandra Hospital, Redditch and Worcester Royal Hospital, Worcester, UK. (294) Ysbyty Gwynedd, Bangor, UK. (295) Yeovil Hospital, Yeovil, UK. (296) University Hospital Hairmyres, East Kilbride, UK. (297) Scunthorpe General Hospital, Scunthorpe, UK. (298) Princess Royal Hospital Brighton, West Sussex, UK. (299) Lincoln County Hospital, Lincoln, UK. (300) Homerton University Hospital, London, UK. (301) Glanagweli General Hospital, Camarthen, UK. (302) Ealing Hospital, Southall, UK. (303) Scarborough General Hospital, Scarborough, UK. (304) Royal Albert Edward Infirmary, Wigan, UK. (305) Queen Elizabeth Hospital Gateshead, Gateshead, UK. (306) North Devon District Hospital, Barnstaple, UK. (307) National Hospital for Neurology and Neurosurgery, London, UK. (308) Eastbourne District General Hospital, East Sussex, UK and Conquest Hospital, East Sussex, UK. (309) Diana Princess of Wales Hospital, Grimsby, UK. (310) The Christie NHS Foundation Trust, Manchester, UK. (311) Prince Philip Hospital, Llanelli, UK. (312) Prince Charles Hospital, Merthyr Tydfil, UK. (313) Golden Jubilee National Hospital, Clydebank, UK. (314) Dorset County Hospital, Dorchester, UK. (315) Calderdale Royal Hospital, Halifax, UK. (316) West Suffolk Hospital, Suffolk, UK. (317) West Cumberland Hospital, Whitehaven, UK. (318) University Hospital Lewisham, London, UK. (319) St John’s Hospital Livingston, Livingston, UK. (320) Sheffield Children’s Hospital, Sheffield, UK. (321) Hinchinbrooke Hospital,
and Compute Canada is acknowledged. TwinsUK is funded by the Wel-
come Trust, Medical Research Council, European Union, the National
Institute for Health Research (NIHR)-funded BioResource, Clinical
Research Facility and Biomedical Research Centre based at Guy’s
and St Thomas’ NHS Foundation Trust in partnership with King’s
College London. These funding agencies had no role in the design,
implementation or interpretation of this study. JBR has served as an
advisor to GlaxoSmithKline and Deerfield Capital. His institution has
received investigator-initiated grant funding from Eli Lilly, GlaxoS-
mithKline and Biogen for projects unrelated to this research. He is
the founder of 5 Prime Sciences. SweCovid received funding from
the European Union’s Horizon 2020 research and innovation program
under Grant agreement No 824110, the SciLifeLab/KAW national
Covid-19 research program project grants to MH (KAW 2020:0182
and KAW 2020:0241) and the Swedish Research Council Grants to
RF (2014-02569 and 2014-07606). Sequencing was performed by the
SNP&SEQ Technology Platform in Uppsala. The facility is part of
the National Genomics Infrastructure (NGI) Sweden and Science for
Life Laboratory. The SNP&SEQ Platform is supported by the Swed-
ish Research Council and the Knut and Alice Wallenberg Foundation.

Declarations

Conflict of interest The authors declare no conflict interests.

Informed consent The patients were informed of this research and
agreed to it through the informed consent process.

Institutional review board statement The study was conducted accord-
ing to the guidelines of the Declaration of Helsinki. The GEN-COVID
is a multicentre academic observational study that was approved by the
Internal Review Boards (IRB) of each participating centre (protocol
code 16917, dated March 16, 2020 for GEN-COVID at the University
Hospital of Siena). BQC19 received ethical approval from the Jewish
General Hospital (JGH) research ethics board (2020-2137). The Swed-
nish part of the study was performed at the general intensive care unit
(ICU) of Uppsala University Hospital, Sweden (a tertiary hospital) and
approved by the National Ethical Review Agency (No. 2020-01623).
The Declaration of Helsinki and its subsequent revisions were fol-
lowed. DeCOI received ethical approval by the Ethical Review Board
(ERB) of the participating hospitals/centres for CoMIRI (Technical
University Munich, Munich, Germany) and BoSCO (Medical Faculty
Bonn, Bonn, Germany; Medical Board of the Saarland, Germany;
University Duisburg-Essen, Germany; Medical Faculty Dusseldorf,
Düsseldorf, Germany). GenOMICC and ISARIC4C were approved
University Duisburg-Essen, Germany; Medical Faculty Duesseldorf,
Bonn, Bonn, Germany; Medical Board of the Saarland, Germany;
University Munich, Munich, Germany) and BoSCO (Medical Faculty
(ERB) of the participating hospitals/centres for CoMIRI (Technical
University Munich, Munich, Germany) and BoSCO (Medical Faculty
Bonn, Bonn, Germany; Medical Board of the Saarland, Germany;
University Duisburg-Essen, Germany; Medical Faculty Dusseldorf,
Düsseldorf, Germany). GenOMICC and ISARIC4C were approved

Data availability and data sharing statement 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).

Open Access This article is licensed under a Creative Commons Attri-
bution 4.0 International License, which permits use, sharing, adapta-
tion, distribution and reproduction in any medium or format, as long
as you give appropriate credit to the original author(s) and the source,
provide a link to the Creative Commons licence, and indicate if changes
were made. The images or other third party material in this article are
included in the article’s Creative Commons licence, unless indicated
otherwise in a credit line to the material. If material is not included in
the article’s Creative Commons licence and your intended use is not
permitted by statutory regulation or exceeds the permitted use, you will
need to obtain permission directly from the copyright holder. To view a
copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Baldassarri M, Picciotti N, Fava F et al (2021a) Shorter androgen
receptor polyQ alleles protect against life-threatening COVID-
19 disease in European males. EBioMedicine 65:103246
Baldassarri M, Fava F, Fallerini C et al (2021b) Severe COVID-19
in hospitalized carriers of single CFTR pathogenic variants. J
Person Med 11(6):558
Bayati A, Kumar R, Francis V, McPherson PS (2021) SARS-CoV-2
infected cells after viral entry via clathrin-mediated endocytosis.
J Biol Chem 296:100306
Benetti E, Gibilerti A, Emiliozzi A et al (2020a) Clinical and molec-
ular characterization of COVID-19 hospitalized patients. PLoS
ONE 15(11):e0242534
Benetti E, Tita R, Spiga O et al (2020b) ACE2 gene variants
may underlie interindividual variability and susceptibility
to COVID-19 in the Italian population. Eur J Hum Genet
28(11):1602–1614
Chen N, Zhou M, Dong X et al (2020) Epidemiological and clinical
characteristics of 99 cases of 2019 novel coronavirus pneumonia
in Wuhan, China: a descriptive study. Lancet 395:507–551
COVID-19 Host Genetics Initiative. https:// www.covid19hg.org/.
Accessed 13 Apr 2020
COVID-19 Host Genetics Initiative et al (2021) Mapping the human
genetic architecture of COVID-19. Published online ahead of
print. Nature. https://doi.org/10.1038/s41586-021-03767-x
COVID-19 Therapeutic Trial Synopsis (2020) WHO R&D blueprint
novel coronavirus. Covid 19 therapeutic trial synopsis. R&D
Blueprint
Covid19hg.org (2021) June 14, 2021 meeting. https://www.covid
19hg.org/blog/2021-06-17-june-14-2021-meeting/. Accessed
13 Apr 2020
Croci S, Venneri MA, Mantovani S et al (2021) The polymorphism
L412F in TLR3 inhibits autophagy and is a marker of severe
COVID-19 in males. Autophagy (in press)
Daga S, Fallerini C, Baldassarri M et al (2021) Employing a systematic
approach to biobanking and analyzing clinical and genetic data for
advancing COVID-19 research. Eur J Hum Genet 29(5):745–759
Danczek P, Bonfield JK, Liddle J et al (2021) Twelve years of SAM
tools and BCF tools. GigaScience 10(2):giab00
Darrah RJ, Jacono FJ, Joshi N et al (2019) AGTR2 absence or antagon-
ism prevents cystic fibrosis pulmonary manifestations. J Cyst
Fibros 18(1):127–134
Elhabyan A, Elyaaacoub S, Sanad E et al (2020) The role of host
genetics in susceptibility to severe viral infections in humans
and insights into host genetics of severe COVID-19: a system-
ic review. Virus Res 289:198163
Fallerini C et al (2021a) Association of Toll-like receptor 7 variants
with life-threatening COVID-19 disease in males: findings from
a nested case-control study. eLife 10:e67569
Fallerini C, Daga S, Benetti E et al (2021b) SELP Aasp603Asn and
severe thrombosis in COVID-19 males: implication for anti-
P-selectin monoclonal antibodies treatment. J Hematol Oncol
14(1):123
Gordon DE, Jang GM, Bouhaddou M et al (2020) A SARS-CoV-2 pro-
tein interaction map reveals targets for drug repurposing. Nature
583(7816):459–468. https://doi.org/10.1038/s41586-020-2286-9
International Multiple Sclerosis Genetics Consortium (2018) Low-
frequency and rare-coding variation contributes to multiple
sclerosis risk. Cell 175(6):1679–1687.e7
Islam MR, Hoque MN, Rahman MS et al. (2020) Genome-wide analysis of SARS-CoV-2 virus strains circulating worldwide implicates heterogeneity. Sci Rep 10(1):14004

Kosmicki JA, Horowitz JE, Banerjee N et al. (2021) Pan-ancestry exome-wide association analyses of COVID-19 outcomes in 586,157 individuals. Am J Hum Genet 108(7):1350–1355. https://doi.org/10.1016/j.ajhg.2021.05.017

Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26(5):589–595

Li X, Zhong X, Wang Y et al. (2021) Clinical determinants of the severity of COVID-19: a systematic review and meta-analysis. PLoS ONE 10(5):16025

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

Livingston E, Bucher K (2020) Coronavirus disease 2019 (COVID-19) in Italy. JAMA 323(14):1335

Marouli E, Graff M, Medina-Gomez C et al. (2017) Rare and low-frequency coding variants alter human adult height. Nature 542(7640):186–190

Mars N, Widén E, Kerninon S et al. (2020) (2020) The role of polygenic risk and susceptibility genes in breast cancer over the course of life. Nat Commun 11(1):6383

Merico D, Isserlin R, Bader GD. Visualizing gene-set enrichment results using the cytoscape plug-in enrichment map. In: Cagney G, Emili A (eds) (2011) Network biology. Methods in molecular biology (methods and protocols), vol 781. Humana Press

Mootha VK, Lindgren CM, Eriksson K et al. (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 34(3):267–273

Pairo-Castineira E, Clohisey S, Klaric L et al. (2020) Genetic mechanisms of critical illness in COVID-19. Nature 591(7848):92–98

Pathak GA, Singh K, Miller-Fleming TW et al. (2021) Post-Mendelian genetic analyses identify susceptibility genes underlying COVID-19 hospitalization. Nat Commun 12(1):4569. https://doi.org/10.1038/s41467-021-24824-z

Pedregosa F, Varoquaux G, Gramfort A et al. (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830

Picciotti N, Benetti E, Fallerini C et al. (2021) Post-Mendelian genetic model in COVID-19. Cardiol Cardiov Med (in press)

Purcell S, Neale B, Todd-Brown K et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3):559–575

Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP (2006) GenePattern 2.0. Nat Genet 38:500–501

Severe Covid-19 GWAS Group, Ellinghaus D, Degenhardt F et al. (2020) Genomewide association study of severe covid-19 with respiratory failure. N Engl J Med 383(16):1522–1534

Shannon P, Markiel A, Ozier O et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13(11):2498–2504

Solanich X, Vargas-Parraga G, van der Made CI et al. (2021) Genetic screening for TLR7 variants in young and previously healthy men with severe COVID-19: a case series. Front Immunol 12:719115

Subramanian A, Tamayo P, Mootha VK et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 102(43):15545–15550

Subramanian S, Tus K, Li QZ et al. (2006) A Tlr7 translocation accelerates systemic autoimmunity in murine lupus. PNAS 103(26):9970–9975

Tan Y, Yang T, Liu P et al. (2017) Association of the OAS3 rs1859330 G/A genetic polymorphism with severity of enterovirus-71 infection in Chinese Han children. Arch Virol 162:2305–2313

Van der Made CI, Simons A, Schuurs-Hoeijmakers J et al. (2020) Presence of genetic variants among young men with severe COVID-19. JAMA 324(7):663–673

Virtanen P, Gommers R, Oliphant TE et al. (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17(3):261–272

Zhang X, Tan Y, Ling Y et al. (2020a) Viral and host factors related to the clinical outcome of COVID-19. Nature 583(7816):437–440

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
