An explainable model of host genetic interactions linked to COVID-19 severity

Anthony Onoja
Scuola Normale Superiore
Nicola Picchiotti
University of Siena, Italy
Chiara Fallerini
University of Siena, Italy
Margherita Baldassarri
University of Siena, Italy
Francesca Fava
University of Siena, Italy
Francesca Colombo
)Istituto di Tecnologie Biomediche – Consiglio Nazionale delle Ricerche
Francesca Chiaromonte
Sant'Anna School of Advanced Studies
Alessandra Renieri
University of Siena, Italy
Simone Furini
University of Siena, Italy
Francesco Raimondi (francesco.raimondi@sns.it)
Scuola Normale Superiore, Pisa https://orcid.org/0000-0002-6891-3178

Article

Keywords:

Posted Date: November 10th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1062190/v1

Abstract

We employed a multifaceted computational strategy to identify the genetic factors contributing to increased risk of severe COVID-19 infection from a Whole Exome Sequencing (WES) dataset of a cohort of 2000 Italian patients. We coupled a stratified k-fold screening, to rank variants more associated with severity, with training of multiple supervised classifiers, to predict severity on the basis of screened features. Feature importance analysis from decision-tree models allowed to identify a handful of 16 variants with highest support which, together with age and gender covariates, were found to be most predictive of COVID-19 severity. When tested on a follow-up cohort, our ensemble of models predicted severity with good accuracy (ACC=81.88%; ROC_AUC=96%; MCC=61.55%). Principal Component Analysis (PCA) and clustering of patients on important variants orthogonally identified two groups of individuals with a higher fraction of severe cases. Our model recapitulated a vast literature of emerging molecular mechanisms and genetic factors linked to COVID-19 response and extends previous landmark Genome Wide Association Studies (GWAS). It revealed a network of interplaying genetic signatures converging on established immune system and inflammatory processes linked to viral infection response, such as JAK-STAT, Cytokine, Interleukin, and C-type lectin receptor signaling. It also identified additional processes cross-talking with immune pathways, such as GPCR signalling, which might offer additional opportunities for therapeutic intervention and patient stratification. Publicly available PheWAS datasets revealed that several variants were significantly associated with phenotypic traits (e.g. “Respiratory or thoracic disease”), confirming their link with COVID-19 severity outcome. Taken together, our analysis suggests that curated genetic information can be effectively integrated along with other patient clinical covariates to forecast COVID-19 disease severity and dissect the underlying host genetic mechanisms for personalized medicine treatments.

Introduction

The coronavirus disease 2019 (COVID-19) pandemic, caused by the infection with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), is challenging at an unprecedented level health, economical and societal systems worldwide. The SARS-CoV-2 infection is characterized by a large variation in consequence ranging from asymptomatic to life-threatening conditions such as viral pneumonia and acute respiratory distress syndrome (ARSD). ARSD is caused by an exaggerated host immune response leading to lung injury, which starts at the epithelial–interstitium–endothelial interface with increased vascular permeability and extravasation of immune cells, mostly macrophages, and granulocytes. Infected epithelial cells and debris bind immune cell receptors, triggering the release of inflammatory cytokines (predominantly IL-6, IL-1, and TNF-α) and activating fibroblasts, resulting in a cytokine release syndrome (Marini & Gattinoni, 2020).

Established host risk factors for disease severity, such as increasing age, male gender, and higher body mass index1, do not explain all variability in disease severity observed across individuals. Genetic factors contributing to COVID-19 susceptibility and severity may provide novel biological insights into disease pathogenesis mechanisms, new drug targets as well as new means for patient stratification, also

considered that, despite the recent development of vaccines, treating the disease remains an important goal in the clinics. The first genetic factors described to contribute to COVID-19 severity were rare loss-of-function variants in genes involved in type I interferon (IFN)(Solanich et al., 2021; Fallerini, Daga, Mantovani, et al., 2021; Zhang et al., 2020; Van Der Made et al., 2020; Bastard et al., 2020). At the same time, several GWAS campaigns investigating the contribution of common genetic variation (Ellinghaus et al., 2020; Pairo-Castineira et al., 2020) to COVID-19 have provided robust support for the involvement of various genomic loci associated with COVID-19 severity and susceptibility, with the strongest finding for severity being located on chromosome 3. Up to now, the Italian GEN-COVID Multicenter Study contributed to the identification of rare variants (Baldassarri, Fava, et al., 2021; Fallerini, Daga, Mantovani, et al., 2021) and common polymorphisms (Fallerini, Daga, Benetti, et al., 2021; Croci et al., 2021; Baldassarri, Picchiotti, et al., 2021) associated with COVID-19 severity through the collection of more than two thousand biospecimens and clinical data from SARS-CoV-2-positive individuals (Daga et al., 2021) and whole exome sequencing (WES) analysis. The COVID-19 Host Genetics Initiative (COVID-19 HGI) (https://www.COVID-19hg.org) has recently provided the most comprehensive picture of host genetic factors linked to COVID-19 severity through meta-analyses of tens of studies from 19 countries (COVID-19 Host Genetics Initiative, 2021).

While GWAS studies provide solid foundations of the host genetic factors individually associated with COVID-19 severity, they most often fail to provide an organic picture about their interplay. By learning non-linear patterns from data in a human interpretable fashion, explainable machine learning algorithms might help in understanding the multifactorial nature of the interactions between host genetics and COVID-19, at the same time providing effective tools for risk and severity forecasting.

In 2020, Italian GEN-COVID Multicenter Study started to investigate how the combination of common and rare variants could determine COVID-19 severity in a pilot study including WES data of a first small cohort of hospitalized patients(Benetti et al., 2020). Previous and ongoing efforts entailed machine learning techniques (i.e. LASSO logistic regression models) in combination with a boolean representation of genetic variants to identify the most informative features associated with the severity which were used to compile an Integrated PolyGenic Score for COVID-19 severity predictions (Fallerini, Picchiotti, et al., 2021) (Picchiotti et al., 2021). In this study, we combined variant case-control screening, supervised binary classifiers training, feature importance analysis, and dimensionality reduction techniques with pathway enrichment and phenotype association studies to identify a few dozens of genetic variants contributing to increased risk of severe COVID-19 infection from a Whole Exome Sequencing (WES) dataset of a cohort of Italian patients.

Results

Comparing genetic variation in severe and asymptomatic individuals

We considered the Whole Exome Sequencing (WES) dataset of germline variants from 1982 European descent patients provided by the GEN-COVID Multicenter Study group(Daga et al., 2021). All subjects were classified according to the grading scheme by the World Health Organization (WHO) defined by the following categories: 0=not hospitalized (a- or pauci-symptomatic); 1=hospitalized without respiratory support; 2=hospitalized O2 supplementation; 3=hospitalized CPAP-biPAP; 4=hospitalized intubated; 5=dead. Demographic (sex, age, and ethnicity) and clinical data (family history, pre-existing chronic conditions, and SARS-CoV-2 related symptoms) were also collected (Fig. 1A).

We started our analysis from a total of 1.057M simple variants which were screened to identify mutations associated with severe patients, likely representing risk factors, from those associated with asymptomatic patients, more likely contributing to protection. We grouped severe patients from clinical groups 5, 4, and 3 which were contrasted against the asymptomatic ones, considered as controls (group 0). We further refined the grading classification by retaining only those patients with severity grade matching the prediction from an ordered logistic regression model using age as input feature for sex-stratified patients (see Methods), yielding a total of 841 samples (518 severe, 323 asymptomatic; Fig. 1B). We employed log odds ratio statistics, using an additive model, to screen variants significantly associated with either severe or asymptomatic groups (see Methods). In order to find a robust set of variants to be used as features set for downstream ML and pathways analysis, we performed the screening on the majority portion (training set) of a randomly split dataset (keeping 80% of the samples for training and 20% for testing). To ensure robustness, we repeated the splitting procedure five times, employing a stratified 5-fold cross-validation scheme, by performing the screening on the training set and finally retaining those variants found to be significantly enriched in each of the five splits (Fig. 1D; see Methods). We found on average 1130 variants significantly enriched across the five folds (Table S1).

Genetic variants predict severity through supervised ML classifiers

We embedded the stratified 5-Fold screening within a supervised classifier training procedure (Fig. 1D; see Methods). For each random split of the dataset, the training set (80% of the original dataset) was used both for variant screening and model training, which was then tested on the corresponding held-out portion (20% of the dataset). For each screened random split, we trained distinct models using a stratified 5-Fold Cross-validation (5-Fold CV) grid search to estimate optimal hyperparameters for supervised classifiers training (Fig. 1D; see Methods). Specifically, for each of the five random splits we trained four independent supervised classifier algorithms, i.e. Logistic regression (LogReg), Support Vector Machine (SVC), Random Forest (RF) and Extreme Gradient Boosting (XGBoost), yielding a total of 20 models. Each of them was then finally tested on the corresponding held-out set from the random split and performance assessments were collected. XGBoost was the algorithm that displayed the smallest drops between training and testing accuracies, achieving the best average performance during testing across the five folds (Fig. 2A; Table S2). In more details, the best XGBoost model had the following performances: Precision=77.27%, Recall=83.33%, MCC=46.69%, ROC_AUC=80, Accuracy=75%, F1-score=80.2% (Table S2).

Overall, we found that 3217 unique variants (out of a total of 3258 unique, screened variants), corresponding to 2546 unique genes, had non-zero coefficients in at least one of the five, decision tree-based models (i.e. RF or XGBoost). However, the XGBoost classifier led to a sharper reduction of relevant variants (1086, corresponding to 1049 genes, with non-zero feature importance in at least one model), consisting of a subset of those identified with the RF models. As expected, clinical covariates such as age and gender were found among the features with the highest median of the distribution of importance coefficients collected from XGBoost models(Fig. 2B). Among this shortlist, only 16 variants (and corresponding genes; Table S1) consistently received non-zero coefficients in all decision tree-based models, out of which 9 variants were found to be enriched (Fig. 2B,C). To confirm the predictive performance of these variants, we re-trained the models by considering only this subset of variants, plus age and gender covariates, and we calculated aggregated performances by considering the median of the probabilities outputted by each model for each sample in the testing set (see Methods). While age and gender covariates alone retained high predictive power (ROC_AUC=80%), the addition of these most informative genetic features led to an increase of performances (ROC_AUC=86%, best model ROC_AUC=91%; Fig. 2D; Table S3).

Remarkably, when we finally tested the ensemble of models trained with only these informative variants on a follow-up cohort of 618 individuals (122 asymptomatic, 496 severe; Fig. 1C) we achieved good performances, either at the individual model level or at the ensemble one (Table S4). In fact, when computing aggregated metrics by considering the median of the probability distribution collected from the ensemble of models (Table S4,5; see Methods), the ensemble of models was able to identify severe patients with good accuracy (ACC=81.88%) and with a ROC_AUC=96%, performing considerably better than the ones obtained by training with only covariates or variants (Fig. 2E; TableS4).

Risk and protective genetic factors impinge on modular, interconnected networks underlying distinct biological processes.

We focused our attention on the subset of variants receiving non-zero feature importance in at least one XGBoost model which we functionally analyzed to provide a mechanistic explanation for their interaction with COVID-19 infection. We performed pathway analysis by mapping mutated genes in a functional interaction (FI) network (i.e. Reactome FI network; see Methods). We built a general FI network(Fig. 3B), as well as networks specific for clinical groups, by grouping variants and genes enriched in severe and asymptomatic patients (Fig. 3A). Pathway analysis on group-specific networks revealed patterns of significantly enriched processes in either asymptomatic or severe patients (Fig. 3A).

In severe patients we found significantly enriched processes associated to cardiomyopathies, e.g. *Arrhythmogenic right ventricular cardiomyopathy* (FDR=4.03e^{-05}), *Calcium signalling pathways* (FDR=4.22e^{-02}), extracellular matrix (ECM), e.g. *ECM-receptor interaction (*FDR=9.22e^{-05}), vesicle-

mediated transport, e.g. *Retinoid metabolism and transport* (FDR= 1.48e^{-02}), *RAB GEFs exchange GTP for GDP on RABs* (FDR= 2.04e^{-02}) and *Clathrin-mediated endocytosis* (FDR= 4.22e^{-02}), transcriptional regulation such as *FOXA2 and FOXA3 transcription factor networks* (FDR=1.48e^{-02}), and immune response such as C-type leptin receptors (CLRs) (FDR=5.67e^{-02}) (Fig. 3A; Table S6). Asymptomatic patients were instead characterized by a distinct set of processes, including *Fanconi anemia pathway* (FDR=7.89e^{-04}), DNA repair processes such as *HDR through HRR or SSA* (FDR=4.84e^{-03}), *Hippo signaling pathway* (FDR=1.64e^{-02}), and *Axon guidance mediated by netrin* (FDR=3.81e^{-02}) (Fig. 3A; Table S7).

The general FI network comprised a total of 344 mutated genes and 630 functional interactions, remarking a high degree of interconnection between affected genes, which participate in different, cross-talking biological processes. Cluster analysis on the general FI network revealed distinct modules characterized by the enrichment of specific pathways, and by a variable composition in terms of variants enriched in either severe or asymptomatic patients. Intriguingly, we found out that no cluster exclusively contained variants enriched in severe or asymptomatic patients. In detail, the largest cluster (i.e. Module 1; 43 nodes) encompassed *Fanconi anemia pathway* (FDR= 2.46e^{-07}) and DNA repair processes such as *HDR through HRR or SSA* (FDR= 4.51e^{-06}) or *Homologous recombination* (FDR=1.76e^{-03}) (Fig. 3B). In this cluster, we found that the gene characterized by the variant with the strongest model support (ms) (i.e. fraction of decision-tree based models assigning non-zero feature importance; see Methods) is *MYBBP1A* rs117615621, which is enriched in asymptomatic patients (log odds ratio (lor)= -1.34; pval= 0.0065; ms=90%; Table S1,S8).

The second-largest module (Module 2; 42 nodes) involves genes mediating signal transduction cascades such as those mediated by Ras GTPases, e.g. Rap1 signaling pathway (FDR=1.01e^{-04}) or MAP kinases, e.g. MAPK signaling pathway (FDR=5.95e^{-04}) (Fig. 3B, C). We also found processes more directly linked to the immune and inflammatory response to the virus, such as the JAK-STAT signaling pathway (FDR=1.11e^{-03}), Cytokine-cytokine receptor interaction (FDR=1.92e^{-03}), and Interleukin-6 family signaling (FDR=1.92e^{-03}) (Fig. 3B, C). All these three pathways are participated by the *CNTFR* gene, which codes for the alpha subunit of the receptor for the ciliary neurotrophic factor, and is affected by a novel variant (chr9:34557898:A: T) enriched in severe patients (lor=1.230663067; pval=0.00021727; Table S1). Intriguingly this variant was ranked in the top20 of genes with the highest median importance (Fig. 2B) and received 100% model support (Fig. 2C), indicating that all the decision-tree-based models considered it as important for the classification of severity. Another variant with 100% support affecting a gene within the same cluster is rs150021157, also significantly enriched among severe patients (lor=1.373871841; pval=0.001927211; Table S1,S8), affecting the *PCSK5* gene, a serine endoprotease which processes various proteins including various cytokines, NGF, renin and which has been reported to regulate the viral life cycle (Decroly et al., 1996).

The third-largest module (Module 3; 38 nodes) is characterized by the *Regulation of nuclear SMAD2/3 signaling* pathway (FDR=1.95e^{-03}) as the most enriched pathway, therefore being tightly interconnected with cluster 2. It was previously shown that SARS nucleocapsid proteins interact with *SMAD3* and

modulate TGF-β signaling (Zhao et al., 2008), another pathway significantly enriched in Module 3 (FDR= 0.014). The latter pathway has also been confirmed to drive a chronic immune response in severe COVID-19 (Ferreira-Gomes et al., 2021).

The variant *SMAD3* rs897912452 (lor=-1.16; pval=0.00051) and the novel *ZMIZ1* 10:79307376:-:GGGGGGGGGG (lor= -1.30608171; pval=6.18e^{-05}) have the highest support (ms=90%) and are found enriched in asymptomatic patients. Additionally, the latter gene *ZMIZ1* participates in another significant pathway, *Coregulation of Androgen receptor activity* (FDR= 0.01), which also entails *AR*, which carries several mutations which, depending on the specific genic locus, can be found enriched either in severe or asymptomatic patients with variable support (Figure 3,S1; Table S1,S8).

We found additional interesting, potentially relevant pathways in the remaining modules. Module 4 (33 nodes) contains genes involved in Deubiquitination (FDR=1.15e^{-05}), a process frequently modified by viral infection (Isaacson & Ploegh, 2009), as well as several other pathways mediating innate immune response such as the TNF receptor signaling pathway (FDR= 1.15e^{-05}), C-leptin receptors (FDR= 7.8e^{-05}) and Toll-like receptor cascades (FDR= 4.76e^{-04})(Fig. S2; Table S8). The *PLEC* gene, involved in filaments network by anchoring intermediate filaments to desmosomes or hemidesmosomes via interlinks with microtubules and microfilaments, belongs to this cluster and it's affected by the variant rs140300753 (lor=1.16, pval=0.002881778, ms=100%), which is enriched in severe and received 100% support from decision-tree based models (Table S1). In Module 5 one of the most significantly enriched pathways is Cilium Assembly (FDR= 2.64e^{-04}), which entails *CEP131* affected by the variant rs2659015, which is enriched in asymptomatic patients (lor=-1.92; pval=0.001517767) and which received 100% model support. Intriguingly *CEP131* has been recently found to be significantly regulated by phosphorylation during viral infection (Vanderboom et al., 2021).

In addition to several other immune response-related processes (e.g. MHC class II antigen presentation in Module 5, FDR=7.13e^{-03}; Table S8), in the remaining clusters we found additional processes with high translational and therapeutic potential. For instance, we found several GPCR-signaling instances significantly enriched in Modules 6 (e.g. *G alpha (i) signaling events*, FDR= 3.69e^{-04}) and 8, which exclusively entails GPCR-downstream signaling pathways and where again the *G alpha (i) signaling events* (FDR= 2.56e-09) and *G alpha (q) signaling events* (FDR= 4.83e-08) are the two downstream pathways most significantly over-represented(Fig. S3; Table S8).

We also found that a few genes whose variants have been identified through our pipeline are among the ones carrying top associations to severity as assessed from studies of the COVID-19 HGI (https://app.COVID-19hg.org/variants)(COVID-19 Host Genetics Initiative, 2021). In detail, variants of 9 out of the 43 genes identified from GWAS studies are also present in our list, including: *ABO, ARL17A, ARL17B, DPP9, LRRC37A, LRRC37A2, RAVER1, TMEM65, ZBTB11* (Table S1).

Severe patients tend to cluster together using only more informative variants.

We applied unsupervised clustering and dimensionality reduction techniques (i.e. Principal Component Analysis (PCA)) to group patients based on the genetic distance calculated by considering the most informative variants selected after screening and supervised machine learning procedure. By projecting the patients on the first two PCs followed by *k-means* clustering (see Methods), we detected three groups of patients on the original cohort (Fig. 4A-C). The two largest clusters were separated by PC1. The largest one, 515 patients, was characterized by a majority of severe cases (78% of the total). The second cluster was instead characterized by a prevalence of asymptomatic patients (70%) of the total. Finally, a third small cluster was identified through the combined usage of PC1 and PC2 and it was characterized almost exclusively by severe patients (95% of 24 patients in total). Notably, the severity of this cluster is only partially explained on the basis of either gender (59 % males and 37 % females; Fig. 4C) or age (Fig.S4A). This cluster was characterized by peculiar genetic features, with a smaller number of variants and a neat prevalence of risk over protection factors (Fig. S4B). Remarkably, a total of 7 (out of 9 overall enriched in severe patients) variants with 100% support from XGB models were also found in this cluster (Table S9). Network analysis of the mutated genes in this predominantly severe cluster highlighted several common processes as well as candidates for drug targeting. In particular, several GPCRs (*ADRB2*, *ADRA1*, *GRM6*), ion channels (*GRIN1*, *CACNA1G*), (receptor tyrosine) kinases (*NTRK1*, *CSF1R*, *GAK*) and nuclear hormone receptors (*AR*, *THRB*) participate to this network and can be readily targeted by approved drugs (Fig. 4C; Table S10).

Important variants are associated to disease traits linked to COVID-19 severe phenotypes

To provide further evidence of a functional relationships between our variants and COVID-19 severe phenotypes, we checked available open-access integrative resources (i.e. Open Target Genetics initiative (Ghoussaini et al., 2021)) which aggregate human GWAS and functional genomics data to link between GWAS-associated loci, variants, and likely causal genes. In particular, we considered Phenome Wide Association Study (PheWAS) analysis considering a wide range of diseases and traits to identify the phenotypes associated with our variants (see Methods). Intriguingly, we found that many variants identified through our approach are associated with traits or phenotypes which might be linked with either risk or protection from severe consequences to the viral infection.

For example, by considering variants with non-zero importance in at least one XGB model we found that those enriched in severe patients were 70% of the total associated with the category “respiratory or thoracic diseases” (see Figure 5A). Among the specific traits with strong associations to more supported variants, we found instances such as “Doctor diagnosed emphysema” (*ITPKA*, rs41277684; *LTK*, rs35932273), the latter variant associated also to “Other alveolar and parietoalveolar pneumopathy”, “Respiratory disorders in diseases classified elsewhere” (*KCNB1*, rs34467662), “Chronic bronchitis/emphysema” (*C12orf43;HNF1A*, rs11065390; *SLC47A2*, rs34399035), “Acute sinusitis”

(*SHANK2*, rs146204677), “Pleural plaque” (*CFAP74*, rs141833643), “Allergic asthma” (*SYTL2*, rs61740616 and rs35751209), “Symptoms and signs involving the circulatory and respiratory systems” (*PCSK5*, rs150021157) (Fig.5B). Although more weakly associated and supported by our models, we also found several associations with chronic obstructive pulmonary disease (COPD) both in “respiratory or thoracic diseases” and in “infectious disease” categories (Table S11). Other disease categories displaying a net prevalence of phenotypic associations for variants enriched among severe were “immune system disease”, with multiple variants associated with specific traits such as “Autoimmune diseases” “Immunodeficiency with predominantly antibody defects” or “Noninfectious disorders of lymphatic channels”, and “pancreatic disease” (Fig.5A; Table S11).

Two of the variants enriched among severe patients which were found by our models to be invariably relevant for severity classification (i.e. *PCSK5* rs150021157 and PLEC rs140300753) were significantly associated with the “*Abnormalities of breathing*” phenotype (pval=0.0000040 and pval=0.00016, respectively), suggesting that patients carrying these variants might be at higher risk due to pre-existing difficulties of breathing (Fig.S5; Table S11).

Other general categories of traits that might be linked to severe COVID-19, such as “*Cardiovascular disease*” or “Infectious disease” showed similar distributions of associations of risk or mitigation factors (Figure S6). Interestingly other categories, such as “Integumentary system disease” showed instead a prevalence of associations with mitigation factors (Figure S6).

Discussion

In this study, we have set up a multifaceted computational strategy to dissect patient genetic variants which might interplay with the SARS-Cov2 virus to increase the risk of, or to protect from, a severe response to infection.

We integrated into a stratified *k*-fold scheme a pipeline to perform variant features screening followed by machine learning model training and testing to robustly identify variants associated with severe response to COVID-19 infection. Our pipeline allowed a drastic reduction of the initial number of variants by several orders of magnitudes: from an initial set of approximately 1M unique variants derived from WES to 1k variants receiving non-zero feature importance in at least one of the decision-tree based modes. By only considering the variants with full support, i.e. always found to have non-zero feature importance in all the decision-based tree models, we further reduced the pool to only 16 variants. Models retrained with only full-support variants (plus age and gender as covariates) achieved superior performances (median ROC_AUC=86%, best model ROC_AUC=91%). Although models trained with only patients age and gender already showed good performances in severity prediction (median ROC_AUC=80%), confirming the predictive power of these covariates, the increase in performance followed by the inclusion of curated genetic information provides the foundation for integrated tools for COVID-19 severity forecast and patient stratification. When tested on a follow-up cohort of more than 600 our models achieved remarkable performances in identifying severe patients with good accuracy (ACC=81.88% and

ROC_AUC=96%), performing considerably better than the ones obtained by training with only covariates or variants (Fig. 2E; TableS4).

The interpretability of our models allowed us to shed new light on the complex landscape of genetic interactions interplaying with virus genetics to contribute to a severe response to COVID-19 in an Italian cohort. Among the 16 variants with 100% support, only 6 genes (37%) were annotated in the largest pathway knowledgebase, i.e. Reactome (Jassal et al., 2020), suggesting that unannotated variants might modulate the interaction with the virus through yet-to-be-discovered biological mechanisms. Intriguingly, we found that two of these highly supported variants, i.e. chr9:34557898:A:T (*CNTFR*) and rs150021157 (*PCSK5*) interact withinin the second-largest module identified on the interaction network of the genes affected by mutations within our study. This cluster, which is moreover the only one characterized by two fully supported variants, is highly enriched in pathways linked to immune response and inflammation, such as the such as JAK-STAT signaling pathway, Cytokine-cytokine receptor interaction, and Interleukin-6 family signaling.

We found that variants enriched in severe patients are involved in cardiomyopathies processes, supporting the established notion that patients with heart disease or its risk factors are at greater risk of severe consequences following COVID-19 infection, including hospitalization, ventilation, or death (Harrison et al., 2021). Additional processes significantly enriched among severe mutations was ECM, whose importance in mediating the interaction with viral particles have been highlighted by affinity-purification proteomics experiments(Gordon et al., 2020). Recent experiments also confirmed a role for integrins in binding to UV-inactivated viral particles, through which outside-inside signaling is elicited via binding to Gα13 (Simons et al., 2021). Vesicle-mediated transport, such as clathrin-mediated endocytosis, has been shown to mediate a key entry point for SARS (Wang et al., 2008). Moreover, C-type leptin receptors have been shown to engage with the virus inducing robust proinflammatory responses in myeloid cells that correlated with COVID-19 severity (Qiao et al., 2021).

We also found several GPCR signaling instances significantly enriched among the network modules. In particular, one cluster of the network was highly specific for these terms, in particular G_i and G_q signalling processes, which mediate vascular inflammation. In particular, the G_q pathway contributes to regulating calcium signaling, which is one of the most enriched processes in our dataset and which leads to endothelial barrier disruption via adherens junction disassembly (Birch et al., 2021). On the other hand, G_q signaling might also contribute to transactivate JAK-STAT pathway via (ERK)1/2 signaling (Birch et al., 2021), the latter in turn also activated by G_i signaling (Goldsmith & Dhanasekaran, 2007). It has also been recently shown that the C5a–C5aR1 axis, which also signals intracellularly through G_q, plays a key role in the pathophysiology of ARSD associated with COVID-19 by starting and maintaining several inflammatory responses through the recruitment and activation of neutrophils and monocytes (Carvelli et al., 2020). Hence, similarly to what we and others previously described in cancer (Raimondi et al., 2019), genetic factors converging on modulating common GPCR downstream signalling pathways might also

contribute to the onset of the inflammatory response related to COVID-19, at the same time offering new therapeutic intervention options for patients with severe forms of COVID-19.

On the other hand, some of the processes that we found significantly enriched among asymptomatic patients have been previously put in connection to SARS viral infection. For example, members of the machinery for DNA damage response have been shown to interact and affect the response to several DNA and RNA viruses (Lilley et al., 2007) and it has been recently demonstrated that these pathways are also triggered by SARS-CoV2 in vitro cellular models (Victor et al., 2021). The Fanconi anemia pathway is tightly linked to DNA repair processes involving homologous recombination and genome integrity (Michl et al., 2016). We therefore speculate that patients carrying variants on these pathways might differently interact with the virus, modulating a milder response to viral infection.

In general we found multiple, recurrent disease traits associated with the variants identified. The variants rs150021157 and rs140300753, characterized by full support during supervised learning, also provide an example of associations to phenotypes that might play a role in COVID-19 severity, such as “Abnormalities of breathing phenotype”. Some categories show a prevalence of associations with risk factors, such as “respiratory or thoracic disease”, including specific traits such as chronic bronchitis, emphysema or COPD (the latter also found in the “infectious disease” category). Other categories enriched for associations with variants enriched in severe patients are “immune system disorders”, including traits such as immunodeficiency with antibody defects, or “pancreas disease”, including several instances mainly associated to Type 2 diabetes, which is a known risk factor for severe COVID-19 (Onder et al., 2020) and whose molecular connection to cytokine storm inflammatory response has now begun to emerge (Melvin et al., 2021). Taken together, these results further corroborate our analysis.

Our model is complementary to previous and ongoing efforts entailing machine learning techniques (i.e. LASSO logistic regression models) and a boolean representation of genetic variants to identify the most informative features associated to severity to compile an Integrated PolyGenic Score for COVID-19 severity predictions (Fallerini, Picchiotti, et al., 2021)(Picchiotti et al., 2021). While we expect that some of the variants identified in this study might be specific for the Italian population, we believe that our approach could be readily trained on different cohorts to identify additional biomarkers for patient stratification in the clinics. Our capability to understand and forecast the genetic factors contributing to COVID-19 disease severity will certainly benefit from the availability of larger sequencing cohorts, the usage of more advanced methods for case-control associations in WES studies, new methodological advancement in the explainable AI field, as well as on our prior- or data-driven knowledge of biological mechanisms linking genetic variants to disease phenotypes.

Methods

Dataset and Pre-processing

We used the whole-exome sequencing (WES) dataset of 1982 European descent patients collected from the GEN-COVID Multicenter Study group coordinated by the University of Siena (https://clinicaltrials.gov/ct2/show/NCT04549831) (Daga et al., 2021). The WES dataset contained a total of 1.057M unique simple variants. Patients were classified according to the grading scheme by the World Health Organization (WHO). The grading classification contained the following categories: 0=not hospitalized (a- or pauci-symptomatic); 1=hospitalized without respiratory support; 2=hospitalized O2 supplementation; 3=hospitalized CPAP-biPAP; 4= hospitalized intubated; 5=dead. We considered patients from more severe groups, i.e. 3,4, and 5, as cases, and asymptomatic patients from group 0, as controls, for a total of 1078 patients. We further refined the grading classification based on an ordinal logistic model which uses age as input feature for sex-stratified patients (Fallerini, Picchiotti, et al., 2021) and we retained only those patients whose grading classification was concordant with the one adjusted by age. This yielded a final set of 841 samples for downstream analysis.

Stratified K-fold split of sample cohort into train and test sets

We embedded a strategy for variant screening in, *stratified k-fold* cross-validation (using the *StratifiedKFold* function from the *scikit-learn* library) scheme to generate 5 random splits, into a training and testing test, of the original dataset. Each fold was constituted by an 80 % training set which was also employed for variant screening and a 20 % testing set. The variants in the test set were curated from the variants screened in the training set. Through the stratified 5-fold approach, we made sure that all the samples of the dataset were employed for testing.

Variant screening

We employed a Log-Odds Ratio (LOR) statistics to perform case-control association and to screen variants associated with either severe or asymptomatic patients in each of the training sets for each of the five folds generated.

GATK best-practices were used to define the variant calling pipeline, as previously described (Fallerini, Picchiotti, et al., 2021). Then, a contingency table to measure the enrichment of reference (*Ref*) or alternative (*Alt*) alleles in either severe or control groups was defined by employing an additive model, whereby homozygous genotype (1/1) has twice the risk (or protection) of the heterozygous type (0/1 or 1/0). We employed the *Table2x2* function from the *statsmodels* library to calculate LORs values and associated p-values and confidence intervals from the the contingency table in Figure S7, respectively employing the functions *log_oddsratio*, *log_oddsratio_pvalue()* and *log_oddsratio_confint()*. We filtered variants with the following characteristics: $p - valuecript >$ and $|LOR| \geq 1$. Variants with LOR > 1 are enriched among severe, while those with LOR < -1 are enriched among asymptomatics.

Feature matrix generation

For each split, we generated a feature matrix for the training set by assigning the allele counts of each screened variant for each sample of the training: i.e. 0 for genotype 0/0, 1 for genotypes 1/0 or 0/1, 2 for

genotype 1/1. The feature matrix for the test set was defined by considering only variants identified as significant after screening the training set of the corresponding split and by assigning the allele count of each sample of the test set. We also included as additional features age, which was normalized, and gender, which was binarized by setting males to 0 and females to 1. Severe patients from group “3+4+5” were given the classification label “1”, the asymptomatic patients from group 0 were given the label “0”.

Feature selection: Removal of Multicollinearity

We employed feature selection techniques to further reduce the number of considered features initially screened through the Log-Odds-Ratio statistics. We tried several approaches, including Lasso, ElasticNet and Multicollinearity, in combination with supervised training approaches (see below). After training several classifiers with the variants selected with each of these methods on a smaller cohort of 1200 samples (*data not shown*), we found that removing multicollinearity from features by considering variant allele counts with correlation coefficients corr.≤|0.8|) gave the best results. The screened features with little or no effects of multicollinearity formed the final 80 % training sets in each fold and the final 20 % corresponding validation sets used for training the supervised machine learning models.

Supervised Binary Classification

We trained supervised learning models for binary classification tasks by employing several algorithms, i.e. Support Vector Machine, Logistic Regression, Random Forest, and Extreme Gradient Boosting classifiers.

Support Vector Classifier (SVC): a popular machine learning method that classifies data points utilizing the concept of hyper-plan and kernel tricks to find fits that best separate the data cloud. In this study, we used the popular Jupyter notebook and *scikit-learn* python package to import the *“sklearn.svm”* SVC classifier model. We first set the SVC default regularization parameter *“C”* to 1, the class weight to “balanced” in order to account for imbalanced classification problems in the dataset. The default linear kernel was used first with the prediction probability set to true. The *GridSearchCV* was used to select the best hyperparameter values for the estimator *“C”, “gamma”*, and the kernel (Linear, Radial Basis Function (RBF), and polynomial) that are critical to the performance of the SVC classifier. The best *GridSearchCV* estimator hyperparameter values that were used to train our dataset were identified as the RBF kernel, *C = 10*, and *gamma* set to *0.1*.

Logistic Regression: a binary classification regression model that uses the logistic function to estimate the parameters of the logistic model. We import from the *scikit-learn* package the “*sklearn.linear_model”* the Logistic Regression model function. We first set the default logistic model classifier parameters; *“class weight = balanced”*, *C = 0.3* and *solver = sag*. The best *GirdSearchCV* estimator values used to train our dataset uses the regularization penalty of l1 (Lasso), *C = 0.7*, and *solver = saga*.

Random Forest(RF): an ensemble learning method that employs a bagging strategy. Multiple decision trees are trained using the same learning algorithm, and then predictions are aggregated from the individual decision tree. From the *“sklearn.ensemble”* library, we import the Random Forest Classifier

function. The RF default model parameters use a class weight set to “balanced”, maximum depth (*max_depth*) of the decision trees was set to 80, the number of features (*max_features*) was set to 2, minimum samples (*min_samples_leaf*) leaf of 3, minimum samples split (*min_samples_split*) of 10, and the number of trees (n_estimators) in the forest was set to 300. The *GridSearchCV* best model estimator parameters were *“bootstrap = True”*, *“max_depth” = 110, “max_features” = 2, “min_samples_leaf” = 5, “min_samples_split” = 10, and “n_estimators” = 100*.

Extreme Gradient Boosted Trees classifier (XGBoost): an ensemble learning classifier family that utilizes boosting strategy to combine a set of weak learners and delivers improved prediction accuracy. We import from the XGBoost package *“xgboost”* library and xgboost function. We defined the data matrix (training feature set and classification label). We set the default XGBoost classifier model parameters class weight to “balanced”, learning objective to *“binary logistic”*. The best *GridSearchCV* estimator parameters values we used to train the dataset were *“learning_rate” = 0.01*, *“max_depth” = 3, “n_estimators” = 140*.

In summary, for each of the four ML models, we performed a parameter optimization through grid search (*GridSearchCV*), using the *accuracy_score* during grid search as a scoring method. We performed a 5-folds cross-validation, by splitting 80% for training and 20% for validation in each fold, repeated three times, using the *StratifiedKFold* function with *n_splits*=5 and *n_repeats*=3. We also set the class weight parameter to “*balanced*" in each of the ML algorithms employed. Both model training and hyperparameters optimization was done with a Python Jupyter notebook interactive web-based development environment using the scikit-learn and the xgboost packages. Model performances on the testing set were evaluated through the following metrics: Accuracy, F1, Precision, Recall, Matthew correlation coefficient (MCC), ROC_AUC.

A consensus voting approach was used to aggregate validation prediction probability scores of the four ML algorithms (SVC, Logistic Regression, Random Forest, and XGBoost classifiers) from each of the (20%) testing sets from each fold by considering the median of the probability distribution collected from the ensemble of models. The features (variants) that received non-zero weight during training of the supervised ML methods (Random Forest and XGBoost classifiers) in each fold were combined across the 5-fold for further interpretation.

We performed a randomization test (i.e. Salzberg’s test) to assess over-fitting (Salzberg, 1997), where we replace the original phenotypic labels of the training matrix with randomly assigned labels while preserving the ratio of the number of positive (severe) and negative (asymptomatic) patients (Table S12).

Feature importance scores

The feature importance assigns weight scores to individual features that interact to predict a particular event in the model. Feature importances for RandomForest and XGBoost models were calculated as the mean decrease in impurity for the feature using the feature importances function from *xgboost*. The

feature importance (weights) scores assigned from these models' predictions were aggregated across the 5-folds to generate a non-zero panel of variants for further downstream analysis.

Final testing on a follow-up cohort

We tested the best performing models trained using most supported variants with and without covariates on a followup cohort of sequenced, Italian patients. An initial set of 838 samples corresponding to grading groups 0, 3, 4 and 5 were refined by applying the same ordered logistic regression classification *adjusted_by_age*, which yielded a final set of 618 individuals (122 asymptomatic, 496 severe). We curated the allele counts of the 16 most informative variants, identified in the first stage of the analysis and model training, from this new set of patients and we used them, together with age and gender, as features for the testing. We evaluated the performances of the ensemble of the 20 models both on an individual as well as on an aggregated level, by calculating aggregated metrics obtained from the median of the probability distribution outputted by the ensemble of the 20 models on the testing samples.

Principal Component Analysis (PCA) and clustering

The variants with non-zero weights from best performing decision-tree models were remapped back into the feature space to form a new feature count matrix covering 100% of the samples (i.e. 841 individuals). This reduced feature matrix was analyzed using Principal Component Analysis (PCA) techniques to reduce the dimensional space. In order for us to do this, we utilized the "*sklearn.decomposition"* library to import the PCA function. We standardized the feature count matrix using the "*sklearn.preprocessing"* library to import the Standard Scaler function. We transformed the normal feature count matrix considering the 1st and 2nd PCA components. We further employ the K-means clustering technique (using the "*sklearn.cluster"* library to import the "*KMeans"* function) to visualize and cluster the 2D PCA components (1st and 2nd dimensions). We set the default cluster size to 3, the maximum iteration (max_iter=1000), and a tolerance value (tol=1e-04). Clusters of patients that express interesting severity patterns were further analyzed using the pathway enrichment for biological interpretations and implications.

Pathway Enrichment Analysis

The pathway enrichment analysis was done using the ReactomeFIViz plugin (Wu et al., 2014) available in Cytoscape (Lotia et al., 2013). The genes corresponding to variants with non-zero feature importance from XGBoost were used to construct a Functional Interaction (FI) network. The general FI network comprised all the genes affected by variants with non-zero feature importances in both patient groups. Node diameter is proportional to the number of variants with non-zero coefficients in any decision-tree-based models. Node color is instead proportional to the LOR with the highest absolute value among the variants associated with a given gene. Modules within the network were identified through spectral partition clustering (Newman, 2006). Reactome pathways over-representation analysis (FDR<0.1) was calculated on either the whole network or for each individual module. We also generated group-specific networks by keeping separated genes with variants enriched in severity from those enriched in asymptomatic and performed pathway over-representation analysis (FDR < 0.1) on the distinct networks.

Retrieving associations between variants and disease traits or phenotypes

We retrieved associations among the variants identified in our study and disease traits or phenotypes through the Open Targets Genetics platform (Ghoussaini et al., 2021). We interrogated the database using the GraphQL query language embedded in a python script and by inputting the variant coordinates (given by chromosome nr, position, Ref, and Alt allele). For each PheWAS association, we retrieved the following data: *eaf, beta, se, nTotal, nCases, oddsRatio, studyId* and *pval*. Only PheWAS with *oddsRatio* > 1 and *pval* <0.001 were considered. The statistics were done only for the variants with non-zero feature importance from XGBoost models.

All the analyses were performed using a customized Python script, with the following libraries: *scipy 1.2.0*, *numpy 1.19.4*, *scikit-learn 0.23.2.*, *statsmodels 0.11.0* and *matplotlib 3.2.1.*

All the scripts and models generated are available at the following URL: https://github.com/Donmaston09/An-explainable-model-of-host-genetic-interactions-linked-to-COVID-19-severity

Declarations

Acknowledgements

F.R. was supported by the Italian Ministry of University and Research through the Department of excellence “Faculty of Sciences” of Scuola Normale Superiore. We gratefully acknowledge computational resources of the Center for High Performance Computing (CHPC) at SNS. We are grateful to Luigi Ambrosio for his initiative and for helpful discussions. 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.COVID-19hg.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 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’.

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

Francesca Mari4,5,6, Sergio Daga4,5, Elisa Benetti4, Mirella Bruttini4,5,6, Maria Palmieri4,5, Susanna Croci4,5, Sara Amitrano6, Katia Capitani4,5, Ilaria Meloni4,5, Elisa Frullanti4,5 , Gabriella Doddato4,5, Mirjam Lista4,5, Giada Beligni4,5, Floriana Valentino4,5, Kristina Zguro4, Rossella Tita6, Annarita Giliberti4,5, Maria Antonietta Mencarelli6, Caterina Lo Rizzo6, Anna Maria Pinto6, Francesca Ariani4,5,6, Laura Di Sarno4,5, Francesca Montagnani4,9, Mario Tumbarello4,9, Ilaria Rancan4,9, Massimiliano Fabbiani9, Barbara Rossetti9, Laura Bergantini10, Miriana D'Alessandro10, Paolo Cameli10, David Bennett10, Federico Anedda11, Simona Marcantonio11, Sabino Scolletta11, Federico Franchi11, Maria Antonietta Mazzei12, Susanna Guerrini12, Edoardo Conticini13, Luca Cantarini13, Bruno Frediani13, Danilo Tacconi14, Chiara Spertilli Raffaelli14, Marco Feri15, Alice Donati15, Raffaele Scala16, Luca Guidelli16, Genni Spargi17, Marta Corridi17, Cesira Nencioni18, Leonardo Croci18, Gian Piero Caldarelli19, Davide Romani20, Paolo Piacentini20 , Maria Bandini20, Elena Desanctis20, Silvia Cappelli20, Anna Canaccini21, Agnese Verzuri21, Valentina Anemoli21, Manola Pisani21, Agostino Ognibene22, Alessandro Pancrazzi22, Maria Lorubbio22, Massimo Vaghi23, Antonella D'Arminio Monforte24, Federica Gaia Miraglia24, Raffaele Bruno25,26, Marco Vecchia25, Massimo Girardis27, Sophie Venturelli27, Stefano Busani27, Andrea Cossarizza28, Andrea Antinori29, Alessandra Vergori29, Arianna Emiliozzi29, Stefano Rusconi30,31, Matteo Siano31, Arianna Gabrieli31, Agostino Riva30,31, Daniela Francisci32, Elisabetta Schiaroli32, Francesco Paciosi32, Andrea Tommasi32, Umberto Zuccon33, Lucia Vietri33, Pier Giorgio Scotton34, Francesca Andretta34, Sandro Panese35, Stefano Baratti35, Renzo Scaggiante36, Francesca Gatti36, Saverio Giuseppe Parisi37, Francesco Castelli38, Eugenia Quiros-Roldan38, Melania Degli Antoni38, Isabella Zanella39,40, Matteo Della Monica41, Carmelo Piscopo41, Mario Capasso42,43,44, Roberta Russo42,43, Immacolata Andolfo42,43, Achille Iolascon42,43, Giuseppe Fiorentino45, Massimo Carella46, Marco Castori46, Filippo Aucella47, Pamela Raggi48, Rita Perna48, Matteo Bassetti49,50, Antonio Di Biagio49,50, Maurizio Sanguinetti51,52, Luca Masucci51,52, Alessandra Guarnaccia51, Serafina Valente53, Oreste De Vivo53, Elena Bargagli10, Marco Mandalà54, Alessia Giorli54, Lorenzo Salerni54, Patrizia Zucchi55, Pierpaolo Parravicini55, Elisabetta Menatti56, Tullio Trotta57, Ferdinando Giannattasio57, Gabriella Coiro57, Fabio Lena58, Gianluca Lacerenza58, Domenico A. Coviello59, Cristina Mussini60, Enrico Martinelli61, Luisa Tavecchia62, Mary Ann Belli62, Lia Crotti63,64,65,66,67, Gianfranco Parati63,64, Maurizio Sanarico68, Filippo Biscarini70, Alessandra Stella70, Marco Rizzi71, Franco Maggiolo71, Diego Ripamonti71, Claudia Suardi72, Tiziana Bachetti73, Maria Teresa La Rovere74, Simona Sarzi-Braga75, Maurizio Bussotti76, Simona Dei78, Sabrina Ravaglia79, Rosangela Artuso80, Elena Andreucci80, Giulia Gori80, Angelica Pagliazzi80, Erika Fiorentini80, Antonio Perrella81, Francesco Bianchi81,4, Paola Bergomi82, Emanuele Catena82, Riccardo Colombo82, Sauro Luchi83, Giovanna Morelli83, Paola Petrocelli83, Sarah Iacopini83, Sara Modica83, Silvia Baroni84, Francesco Vladimiro Segala85, Francesco Menichetti86, Marco Falcone86, Giusy Tiseo86, Chiara Barbieri86, Tommaso Matucci86, Davide Grassi87, Claudio Ferri88, Franco Marinangeli89, Francesco Brancati90, Antonella Vincenti91, Valentina Borgo91, Stefania Lombardi91, Mirco Lenzi91, Massimo Antonio Di Pietro92, Francesca Vichi92, Benedetta Romanin92, Letizia Attala92, Cecilia Costa92, Andrea Gabbuti92, Roberto Menè63,64, Marta Colaneri25, Patrizia Casprini93, Giuseppe Merla94,95, Gabriella Maria Squeo94, Marcello Maffezzoni96, Stefania Mantovani97, Mario U. Mondelli97, Serena Ludovisi98

9. Department of Medical Sciences, Infectious and Tropical Diseases Unit, Azienda Ospedaliera Universitaria Senese, Siena, Italy

10. Unit of Respiratory Diseases and Lung Transplantation, Department of Internal and Specialist Medicine, University of Siena, Italy

11. Department of Emergency and Urgency, Medicine, Surgery and Neurosciences, Unit of Intensive Care Medicine, Siena University Hospital, Italy

12.Department of Medical, Surgical and Neuro Sciences and Radiological Sciences, Unit of Diagnostic Imaging, University of Siena, Italy

13. Rheumatology Unit, Department of Medicine, Surgery and Neurosciences, University of Siena, Policlinico Le Scotte, Italy

14. Department of Specialized and Internal Medicine, Infectious Diseases Unit, San Donato Hospital Arezzo, Italy

15. Department of Emergency, Anesthesia Unit, San Donato Hospital, Arezzo, Italy

16. Department of Specialized and Internal Medicine, Pneumology Unit and UTIP, San Donato Hospital, Arezzo, Italy

17. Department of Emergency, Anesthesia Unit, Misericordia Hospital, Grosseto, Italy

18. Department of Specialized and Internal Medicine, Infectious Diseases Unit, Misericordia Hospital, Grosseto, Italy

19. Clinical Chemical Analysis Laboratory, Misericordia Hospital, Grosseto, Italy

20. Dipartimento di Prevenzione, Azienda USL Toscana Sud Est, Italy

21. Dipartimento Tecnico-Scientifico Territoriale, Azienda USL Toscana Sud Est, Italy

22. Clinical Chemical Analysis Laboratory, San Donato Hospital, Arezzo, Italy

23. Chirurgia Vascolare, Ospedale Maggiore di Crema, Italy

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

25. Division of Infectious Diseases I, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy.

26.Department of Clinical, Surgical, Diagnostic, and Pediatric Sciences, University of Pavia, Pavia, Italy.

27. Department of Anesthesia and Intensive Care, University of Modena and Reggio Emilia, Modena, Italy

28.Department of Medical and Surgical Sciences for Children and Adults, University of Modena and Reggio Emilia, Modena, Italy

29. HIV/AIDS Department, National Institute for Infectious Diseases, IRCCS, Lazzaro Spallanzani, Rome, Italy

30. III Infectious Diseases Unit, ASST-FBF-Sacco, Milan, Italy

31.Department of Biomedical and Clinical Sciences Luigi Sacco, University of Milan, Milan, Italy

32. Infectious Diseases Clinic, "Santa Maria" Hospital, University of Perugia, Perugia, Italy

33.Respiratory Diseases Unit, "Santa Maria degli Angeli" Hospital, Pordenone, Italy

34.Department of Infectious Diseases, Treviso Hospital, Local Health Unit 2 Marca Trevigiana, Treviso, Italy

35. Clinical Infectious Diseases, Mestre Hospital, Venezia, Italy.

36. Infectious Diseases Clinic, ULSS1, Belluno, Italy

37.Department of Molecular Medicine, University of Padova, Italy

38. Department of Infectious and Tropical Diseases, University of Brescia and ASST Spedali Civili Hospital, Brescia, Italy

39. Department of Molecular and Translational Medicine, University of Brescia, Italy;

40.Clinical Chemistry Laboratory, Cytogenetics and Molecular Genetics Section, Diagnostic Department, ASST Spedali Civili di Brescia, Italy

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

42. Department of Molecular Medicine and Medical Biotechnology, University of Naples Federico II, Naples, Italy

43. CEINGE Biotecnologie Avanzate, Naples, Italy

44. IRCCS SDN, Naples, Italy

45. Unit of Respiratory Physiopathology, AORN dei Colli, Monaldi Hospital, Naples, Italy

46. Division of Medical Genetics, Fondazione IRCCS Casa Sollievo della Sofferenza Hospital, San Giovanni Rotondo, Italy

47.Department of Medical Sciences, Fondazione IRCCS Casa Sollievo della Sofferenza Hospital, San Giovanni Rotondo, Italy

48.Clinical Trial Office, Fondazione IRCCS Casa Sollievo della Sofferenza Hospital, San Giovanni Rotondo, Italy

49. Department of Health Sciences, University of Genova, Genova, Italy

50. Infectious Diseases Clinic, Policlinico San Martino Hospital, IRCCS for Cancer Research Genova, Italy

51. Microbiology, Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Catholic University of Medicine, Rome, Italy

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

53.Department of Cardiovascular Diseases, University of Siena, Siena, Italy

54.Otolaryngology Unit, University of Siena, Italy

55.Department of Internal Medicine, ASST Valtellina e Alto Lario, Sondrio, Italy

56. Study Coordinator Oncologia Medica e Ufficio Flussi Sondrio, Italy

57. First Aid Department, Luigi Curto Hospital, Polla, Salerno, Italy

58.Department of Pharmaceutical Medicine, Misericordia Hospital, Grosseto, Italy.

59. U.O.C. Laboratorio di Genetica Umana, IRCCS Istituto G. Gaslini, Genova, Italy

60. Infectious Diseases Clinics, University of Modena and Reggio Emilia, Modena, Italy

61. Department of Respiratory Diseases, Azienda Ospedaliera di Cremona, Cremona, Italy

62. U.O.C. Medicina, ASST Nord Milano, Ospedale Bassini, Cinisello Balsamo (MI), Italy

63.Istituto Auxologico Italiano, IRCCS, Department of Cardiovascular, Neural and Metabolic Sciences, San Luca Hospital, Milan, Italy

64. Department of Medicine and Surgery, University of Milano-Bicocca, Milan, Italy

65. Istituto Auxologico Italiano, IRCCS, Center for Cardiac Arrhythmias of Genetic Origin, Milan, Italy

66. Istituto Auxologico Italiano, IRCCS, Laboratory of Cardiovascular Genetics, Milan, Italy

67. Member of the European Reference Network for Rare, Low Prevalence and Complex Diseases of the Heart-ERN GUARD-Heart

68. Independent Data Scientist, Milan, Italy

69. Scuola Normale Superiore, Pisa, Italy

70.CNR-Consiglio Nazionale delle Ricerche, Istituto di Biologia e Biotecnologia Agraria (IBBA), Milano,Italy

71.Unit of Infectious Diseases, ASST Papa Giovanni XXIII Hospital, Bergamo, Italy

72. Fondazione per la ricerca Ospedale di Bergamo, Bergamo, Italy

73. Direzione Scientifica, Istituti Clinici Scientifici Maugeri IRCCS, Pavia, Italy

74. Istituti Clinici Scientifici Maugeri IRCCS, Department of Cardiology, Institute of Montescano, Pavia, Italy

75. Istituti Clinici Scientifici Maugeri, IRCCS, Department of Cardiac Rehabilitation, Institute of Tradate (VA), Italy

76. Istituti Clinici Scientifici Maugeri IRCCS, Department of Cardiology, Institute of Milan, Milan, Italy

77. Core Research Laboratory, ISPRO, Florence, Italy

78. Health Management, Azienda USL Toscana Sudest, Tuscany, Italy

79. IRCCS C. Mondino Foundation, Pavia, Italy

80. Medical Genetics Unit, Meyer Children's University Hospital, Florence, Italy

81. Department of Medicine, Pneumology Unit, Misericordia Hospital, Grosseto, Italy.

82. Department of Anesthesia and Intensive Care Unit, ASST Fatebenefratelli Sacco, Luigi Sacco Hospital, Polo Universitario, University of Milan, Milan

83. Infectious Disease Unit, Hospital of Lucca, Italy

84. Department of Diagnostic and Laboratory Medicine, Institute of Biochemistry and Clinical Biochemistry, Fondazione Policlinico Universitario A. Gemelli IRCCS, Catholic University of the Sacred Heart, Rome, Italy.

85. Clinic of Infectious Diseases, Catholic University of the Sacred Heart, Rome, Italy

86. Department of Clinical and Experimental Medicine, Infectious Diseases Unit, University of Pisa, Pisa, Italy

87. Department of Clinical Medicine, Public Health, Life and Environment Sciences, University of L'Aquila, L'Aquila, Italy

88.Department of Clinical Medicine, Public Health, Life and Environment Sciences, University of L'Aquila, Italy

89. Anesthesiology and Intensive Care, University of L'Aquila, L'Aquila, Italy

90. Medical Genetics Unit, Department of Life, Health and Environmental Sciences, University of L'Aquila, L'Aquila, Italy

91. Infectious Disease Unit, Hospital of Massa, Italy

92. Infectious Diseases Unit, Santa Maria Annunziata Hospital, USL Centro, Florence, Italy

93. Laboratory of Clinical Pathology and Immunoallergy, Florence-Prato, Italy

94. Laboratory of Regulatory and Functional Genomics, Fondazione IRCCS Casa Sollievo della Sofferenza, San Giovanni Rotondo (Foggia), Italy

95. Department of Molecular Medicine and Medical Biotechnology, University of Naples Federico II, Naples, Italy.

96. University of Pavia, Pavia, Italy

97. Division of Clinical Immunology and Infectious Diseases, Department of Medicine, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy

98 .Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy.

References

Fallerini, C., Daga, S., Benetti, E., Picchiotti, N., Zguro, K., Catapano, F., Baroni, V., Lanini, S., Bucalossi, A., Marotta, G., Colombo, F., Baldassarri, M., Fava, F., Beligni, G., Di Sarno, L., Alaverdian, D., Palmieri, M., Croci, S., Isidori, A. M., … Mari, F. (2021). SELP Asp603Asn and severe thrombosis in COVID-19 males. *Journal of Hematology and Oncology*, *14*(1), 1–4. https://doi.org/10.1186/S13045-021-01136-9/TABLES/1

Melvin, W. J., Audu, C. O., Davis, F. M., Sharma, S. B., Joshi, A., DenDekker, A., Wolf, S., Barrett, E., Mangum, K., Zhou, X., Bame, M., Ruan, A., Obi, A., Kunkel, S. L., Moore, B. B., & Gallagher, K. A. (2021). Coronavirus induces diabetic macrophage-mediated inflammation via SETDB2. *Proceedings of the National Academy of Sciences*, *118*(38). https://doi.org/10.1073/PNAS.2101071118

Fallerini, C., Picchiotti, N., Baldassarri, M., Zguro, K., Daga, S., Fava, F., Benetti, E., Amitrano, S., Bruttini, M., Palmieri, M., Croci, S., Lista, M., Beligni, G., Valentino, F., Meloni, I., Tanfoni, M., Colombo, F., Cabri, E., Fratelli, M., … Furini, S. (2021). Common, low-frequency, rare, and ultra-rare coding variants contribute to COVID-19 severity. *MedRxiv*, *16*, 2021.09.03.21262611. https://doi.org/10.1101/2021.09.03.21262611

Victor, J., Deutsch, J., Whitaker, A., Lamkin, E. N., March, A., Zhou, P., Botten, J. W., & Chatterjee, N. (2021). SARS-CoV-2 triggers DNA damage response in Vero E6 cells. *BioRxiv*, 2021.09.08.459535. https://doi.org/10.1101/2021.09.08.459535

Solanich, X., Vargas-Parra, G., van der Made, C. I., Simons, A., Schuurs-Hoeijmakers, J., Antolí, A., del Valle, J., Rocamora-Blanch, G., Setién, F., Esteller, M., van Reijmersdal, S. V., Riera-Mestre, A., Sabater-Riera, J., Capellá, G., van de Veerdonk, F. L., van der Hoven, B., Corbella, X., Hoischen, A., & Lázaro, C. (2021). Genetic Screening for TLR7 Variants in Young and Previously Healthy Men With Severe COVID-19. *Frontiers in Immunology*, *12*, 2965. https://doi.org/10.3389/FIMMU.2021.719115/BIBTEX

Harrison, S. L., Buckley, B. J. R., Rivera-Caravaca, J. M., Zhang, J., & Lip, G. Y. H. (2021). Cardiovascular risk factors, cardiovascular disease, and COVID-19: an umbrella review of systematic reviews. *European Heart Journal - Quality of Care and Clinical Outcomes*, *7*(4), 330–339. https://doi.org/10.1093/EHJQCCO/QCAB029

Simons, P., Rinaldi, D. A., Bondu, V., Kell, A. M., Bradfute, S., Lidke, D., & Buranda, T. (2021). Integrin activation is an essential component of SARS-CoV-2 infection. *BioRxiv*, 2021.07.20.453118. https://doi.org/10.1101/2021.07.20.453118

COVID-19 Host Genetics Initiative. (2021). Mapping the human genetic architecture of COVID-19. *Nature 2021*, 1–8. https://doi.org/10.1038/s41586-021-03767-x

Baldassarri, M., Fava, F., Fallerini, C., Daga, S., Benetti, E., Zguro, K., Amitrano, S., Valentino, F., Doddato, G., Giliberti, A., Di Sarno, L., Palmieri, M., Carriero, M. L., Alaverdian, D., Beligni, G., Iuso, N., Castelli, F., Quiros-roldan, E., Mondelli, M. U., … Gabbi, C. (2021). Severe COVID-19 in Hospitalized Carriers of Single CFTR Pathogenic Variants. *Journal of Personalized Medicine 2021, Vol. 11, Page 558*, *11*(6), 558. https://doi.org/10.3390/JPM11060558

Qiao, L., J, L., S, Z., MF, G. C., M, L.-R., J, D., X, R., P, D.-Y., H, T., T, K., J, S., ME, K., Z, Z., ST, Y., BT, M., RE, C., F, T., X, R., X, C., … J, W. (2021). SARS-CoV-2 exacerbates proinflammatory responses in myeloid cells through C-type lectin receptors and Tweety family member 2. *Immunity*, *54*(6), 1304-1319.e9. https://doi.org/10.1016/J.IMMUNI.2021.05.006

Ferreira-Gomes, M., Kruglov, A., Durek, P., Heinrich, F., Tizian, C., Heinz, G. A., Pascual-Reguant, A., Du, W., Mothes, R., Fan, C., Frischbutter, S., Habenicht, K., Budzinski, L., Ninnemann, J., Jani, P. K., Guerra, G. M., Lehmann, K., Matz, M., Ostendorf, L., … Mashreghi, M.-F. (2021). SARS-CoV-2 in severe COVID-19 induces a TGF-β-dominated chronic immune response that does not target itself. *Nature Communications 2021 12:1*, *12*(1), 1–14. https://doi.org/10.1038/s41467-021-22210-3

Croci, S., Venneri, M. A., Mantovani, S., Fallerini, C., Benetti, E., Picchiotti, N., Campolo, F., Imperatore, F., Palmieri, M., Daga, S., Gabbi, C., Montagnani, F., Beligni, G., Farias, T. D. J., Carriero, M. L., Sarno, L. D., Alaverdian, D., Aslaksen, S., Cubellis, M. V., … Squeo, G. M. (2021). The polymorphism L412F in TLR3

inhibits autophagy and is a marker of severe COVID-19 in males. *MedRxiv*, 2021.03.23.21254158. https://doi.org/10.1101/2021.03.23.21254158

Baldassarri, M., Picchiotti, N., Fava, F., Fallerini, C., Benetti, E., Daga, S., Valentino, F., Doddato, G., Furini, S., Giliberti, A., Tita, R., Amitrano, S., Bruttini, M., Croci, S., Meloni, I., Pinto, A. M., Iuso, N., Gabbi, C., Sciarra, F., … Gut, M. (2021). Shorter androgen receptor polyQ alleles protect against life-threatening COVID-19 disease in European males. *EBioMedicine*, *65*, 103246. https://doi.org/10.1016/J.EBIOM.2021.103246/ATTACHMENT/603FE86C-BFBF-429A-A091-CD6556D2B2FA/MMC6.XLSX

Fallerini, C., Daga, S., Mantovani, S., Benetti, E., Picchiotti, N., Francisci, D., Paciosi, F., Schiaroli, E., Baldassarri, M., Fava, F., Palmieri, M., Ludovisi, S., Castelli, F., Quiros-Roldan, E., Vaghi, M., Rusconi, S., Siano, M., Bandini, M., Spiga, O., … Dei, S. (2021). Association of toll-like receptor 7 variants with life-threatening COVID-19 disease in males: Findings from a nested case-control study. *ELife*, *10*. https://doi.org/10.7554/ELIFE.67569

Birch, C. A., Molinar-Inglis, O., & Trejo, J. (2021). Subcellular hot spots of GPCR signaling promote vascular inflammation. *Current Opinion in Endocrine and Metabolic Research*, *16*, 37. https://doi.org/10.1016/J.COEMR.2020.07.011

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, A. M., Raimondi, F., Stella, A., Biscarini, F., Picchiotti, N., Gori, M., … Frullanti, E. (2021). Employing a systematic approach to biobanking and analyzing clinical and genetic data for advancing COVID-19 research. *European Journal of Human Genetics*, 1–15. https://doi.org/10.1038/s41431-020-00793-7

Vanderboom, P. M., Mun, D.-G., Madugundu, A. K., Mangalaparthi, K. K., Saraswat, M., Garapati, K., Chakraborty, R., Ebihara, H., Sun, J., & Pandey, A. (2021). Proteomic Signature of Host Response to SARS-CoV-2 Infection in the Nasopharynx. *Molecular & Cellular Proteomics*, *20*, 100134. https://doi.org/10.1016/J.MCPRO.2021.100134

Picchiotti, N., Benetti, E., Fallerini, C., Daga, S., Giliberti, A., Tita, R., Amitrano, S., Bruttini, M., Sarno, D., Iuso, N., Alaverdian, D., Beligni, G., Croci, S., Crawley, F. P., Frullanti, E., & Mari, F. (2021). *Post-Mendelian genetic model in COVID-19*.

Zhang, Q., Bastard, P., Liu, Z., Pen, J. Le, Moncada-Velez, M., Chen, J., Ogishi, M., Sabli, I. K. D., Hodeib, S., Korol, C., Rosain, J., Bilguvar, K., Ye, J., Bolze, A., Bigio, B., Yang, R., Arias, A. A., Zhou, Q., Zhang, Y., … Casanova, J.-L. (2020). Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. *Science (New York, N.Y.)*, *370*(6515). https://doi.org/10.1126/SCIENCE.ABD4570

Van Der Made, C. I., Simons, A., Schuurs-Hoeijmakers, J., Van Den Heuvel, G., Mantere, T., Kersten, S., Van Deuren, R. C., Steehouwer, M., Van Reijmersdal, S. V., 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., … Hoischen, A. (2020). Presence of Genetic Variants Among Young Men With Severe COVID-19. *JAMA*, *324*(7), 663–673. https://doi.org/10.1001/JAMA.2020.13719

Carvelli, J., Demaria, O., Vély, F., Batista, L., Chouaki Benmansour, N., Fares, J., Carpentier, S., Thibult, M.-L., Morel, A., Remark, R., André, P., Represa, A., Piperoglou, C., Cordier, P. Y., Le Dault, E., Guervilly, C., Simeone, P., Gainnier, M., Morel, Y., … Vivier, E. (2020). Association of COVID-19 inflammation with activation of the C5a–C5aR1 axis. *Nature 2020 588:7836*, *588*(7836), 146–150. https://doi.org/10.1038/s41586-020-2600-6

Benetti, E., Tita, R., Spiga, O., Ciolfi, A., Birolo, G., Bruselles, A., Doddato, G., Giliberti, A., Marconi, C., Musacchia, F., Pippucci, T., Torella, A., Trezza, A., Valentino, F., Baldassarri, M., Brusco, A., Asselta, R., Bruttini, M., Furini, S., … Pinto, A. M. (2020). ACE2 gene variants may underlie interindividual variability and susceptibility to COVID-19 in the Italian population. *European Journal of Human Genetics 2020 28:11*, *28*(11), 1602–1614. https://doi.org/10.1038/s41431-020-0691-z

Marini, J. J., & Gattinoni, L. (2020). Management of COVID-19 Respiratory Distress. *JAMA*, *323*(22), 2329–2330. https://doi.org/10.1001/JAMA.2020.6825

Onder, G., Rezza, G., & Brusaferro, S. (2020). Case-Fatality Rate and Characteristics of Patients Dying in Relation to COVID-19 in Italy. *JAMA*, *323*(18), 1775–1776. https://doi.org/10.1001/JAMA.2020.4683

Gordon, D. E., Jang, G. M., Bouhaddou, M., Xu, J., Obernier, K., White, K. M., O'Meara, M. J., Rezelj, V. V., Guo, J. Z., Swaney, D. L., Tummino, T. A., Hüttenhain, R., Kaake, R. M., Richards, A. L., Tutuncuoglu, B., Foussard, H., Batra, J., Haas, K., Modak, M., … Krogan, N. J. (2020). A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature 2020 583:7816*, *583*(7816), 459–468. https://doi.org/10.1038/s41586-020-2286-9

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., Loney, F., May, B., Milacic, M., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Weiser, J., … D'Eustachio, P. (2020). The reactome pathway knowledgebase. *Nucleic Acids Research*, *48*(D1), D498–D503. https://doi.org/10.1093/nar/gkz1031

Bastard, P., Rosen, L. B., Zhang, Q., Zhang, Y., Dorgham, K., Béziat, V., Puel, A., Lorenzo, L., Bizien, L., Assant, S., Fillipot, Q., Seeleuthner, Y., Hadjadj, J., Bigio, B., Michael, S., Shaw, E., Chauvin, S. D., Belot, A., & Rieux-laucat, F. (2020). IgG autoantibodies against type I IFNs in patients with severe COVID-19. *Science*, *4585*(September), 1–19.

Ellinghaus, D., Degenhardt, F., Bujanda, L., Buti, M., Albillos, A., Invernizzi, P., Fernández, J., Prati, D., Baselli, G., Asselta, R., Grimsrud, M. M., Milani, C., Aziz, F., Kässens, J., May, S., Wendorff, M., Wienbrandt, L., Uellendahl-Werth, F., Zheng, T., … Karlsen, T. H. (2020). Genomewide Association Study of Severe COVID-19 with Respiratory Failure. *The New England Journal of Medicine*. https://doi.org/10.1056/NEJMoa2020283

Pairo-Castineira, E., Clohisey, S., Klaric, L., Bretherick, A., Rawlik, K., Pasko, D., Walker, S., Richmond, A., Head Fourman, M., Russell, C. D., Law, A., Furniss, J., Gountouna, E., Wrobel, N., Moutsianas, L., Wang, B., Meynert, A., Yang, Z., Zhai, R., … Scott, R. (2020). Genetic mechanisms of critical illness in COVID-19. In *medRxiv* (Vol. 17, Issue 8). https://doi.org/10.1101/2020.09.24.20200048

Raimondi, F., Inoue, A., Kadji, F. M. N., Shuai, N., Gonzalez, J.-C., Singh, G., de la Vega, A. A., Sotillo, R., Fischer, B., Aoki, J., & others. (2019). Rare, functional, somatic variants in gene families linked to cancer genes: GPCR signaling as a paradigm. *Oncogene*, *38*(38), 6491–6506.

Michl, J., Zimmer, J., & Tarsounas, M. (2016). Interplay between Fanconi anemia and homologous recombination pathways in genome integrity. *The EMBO Journal*, *35*(9), 909–923. https://doi.org/10.15252/EMBJ.201693860

Isaacson, M. K., & Ploegh, H. L. (2009). Ubiquitination, Ubiquitin-like Modifiers, and Deubiquitination in Viral Infection. *Cell Host & Microbe*, *5*(6), 559–570. https://doi.org/10.1016/J.CHOM.2009.05.012

Zhao, X., Nicholls, J. M., & Chen, Y. G. (2008). Severe Acute Respiratory Syndrome-associated Coronavirus Nucleocapsid Protein Interacts with Smad3 and Modulates Transforming Growth Factor-β Signaling. *Journal of Biological Chemistry*, *283*(6), 3272–3280. https://doi.org/10.1074/JBC.M708033200

Wang, H., Yang, P., Liu, K., Guo, F., Zhang, Y., Zhang, G., & Jiang, C. (2008). SARS coronavirus entry into host cells through a novel clathrin- and caveolae-independent endocytic pathway. *Cell Research 2008 18:2*, *18*(2), 290–301. https://doi.org/10.1038/cr.2008.15

Goldsmith, Z., & Dhanasekaran, D. (2007). G protein regulation of MAPK networks. *Oncogene*, *26*(22), 3122–3142. https://doi.org/10.1038/SJ.ONC.1210407

Lilley, C., Schwartz, R., & Weitzman, M. (2007). Using or abusing: viruses and the cellular DNA damage response. *Trends in Microbiology*, *15*(3), 119–126. https://doi.org/10.1016/J.TIM.2007.01.003

Decroly, E., Wouters, S., Bello, C. Di, Lazure, C., Ruysschaert, J. M., & Seidah, N. G. (1996). Identification of the Paired Basic Convertases Implicated in HIV gp160 Processing Based on in Vitro Assays and Expression in CD4+ Cell Lines. *Journal of Biological Chemistry*, *271*(48), 30442–30450. https://doi.org/10.1074/JBC.271.48.30442

Supplemental Legends

Supplementary Figure Legends

Supplementary Figure1: top) Reactome FI network of genes of module 3 affected by variants with non-zero feature importance from XGBoost. Node dimension and coloring is same as in Fig; bottom) barchart of the enriched processes within the module.

Supplementary Figure2: top) Reactome FI network of genes of module 4 affected by variants with non-zero feature importance from XGBoost. Node dimension and coloring is same as in Fig; bottom) barchart of the enriched processes within the module.

Supplementary Figure3: top) Reactome FI network of genes of module 8 affected by variants with non-zero feature importance from XGBoost. Node dimension and coloring is same as in Fig; bottom) barchart of the enriched processes within the module.

Supplementary Figure4: A) age distribution of the patients in the three clusters identified by PCA and k-means clustering considering non-zero importance variants in the 2000 patients cohort; B) variant distribution in the three cluster.

Supplementary Figure5: scatter plot showing variant-specific traits associated within the “Phenotype” category. Dot diameter and color is as in Fig. 5.

Supplementary Figure6: barchart of the associations of traits to variants enriched in either severe (red) or asymptomatic (blue) patients for general categories.

Supplementary Figure7: contingency table employed to perform log-odds ratio statistics for case(severe)-control(asymptomatic) associations.

Supplementary Table Legends

Supplementary Table1: screened variants annotated with log-odds ratio statistics, statistics of the importance from decision-tree models, dbSNP ids and Reactome pathways and clusters

Supplementary Table2: performances of training and testing of the predictors trained with screened variants on the 5 Folds plus age and gender covariates

Supplementary Table3: performances of training and testing the classifier with 100% supported variants on the 5 Folds with and without covariates

Supplementary Table4: performances of the models trained with 16 fully supported variants, with and without covariates, tested in a followup cohort of 618 sequenced patients

Supplementary Table5: patients level prediction probabilities of the models trained with 16 fully supported variants and covariates and tested in a followup cohort of 618 sequenced patients. Probabilities of individual models as well as aggregated (median) over folds and algorithm are shown.

Supplementary Table6: significantly over-represented pathways from ReactomeFI network builts with variants enriched in severe patients

Supplementary Table7: significantly over-represented pathways from ReactomeFI network built with variants enriched in asymptomatic patients

Supplementary Table8: significantly over-represented pathways in the module of the ReactomeFI general network

Supplementary Table9: genes with variants affecting the patients in the cluster with highest fraction of severe cases

Supplementary Table10: approved drug-target interactions from Reactome FI network of mutated genes in the most severe cluster

Supplementary Table11: significant (pval < 0.001) PheWAS associations between disease traits and important variants

Supplementary Table12: Salzberg's test on training performances

Figures

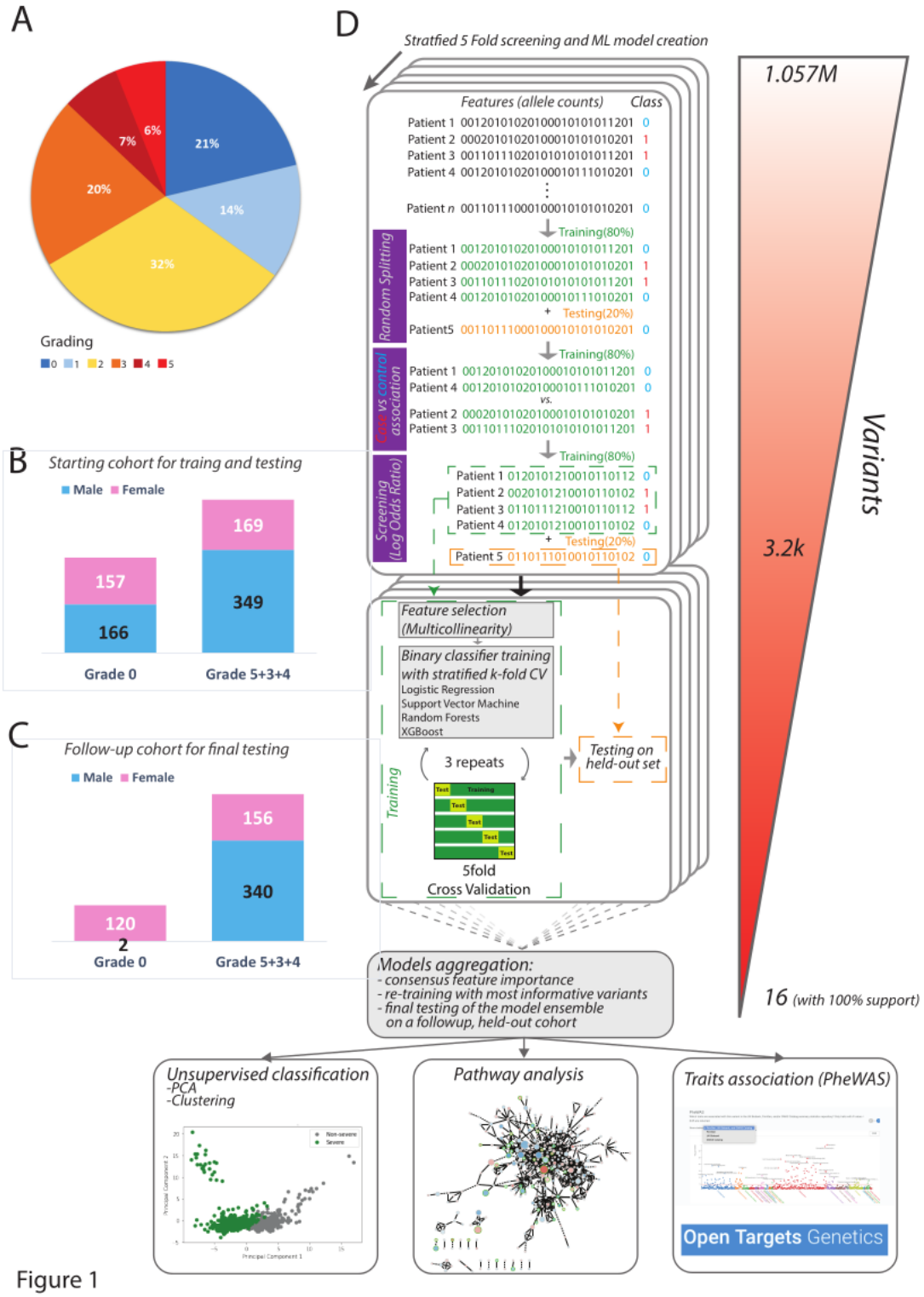

Figure 1

Figure 1

A) piechart with the fraction of sequenced patients for each grading group; B) stacked bar-charts with distribution of patients in the two groups (severe=5+4+3; asymptomatic=0), and their gender composition, whose variants were used for screening, training and initial testing; C) stacked bar-charts with distribution of patients in the two groups (severe=5+4+3; asymptomatic=0), and their gender

composition, from a follow-up cohort used for final testing of the model; D) workflow of the bottom-up computational strategy to identify and interpret variants linked to COVID-19 severity

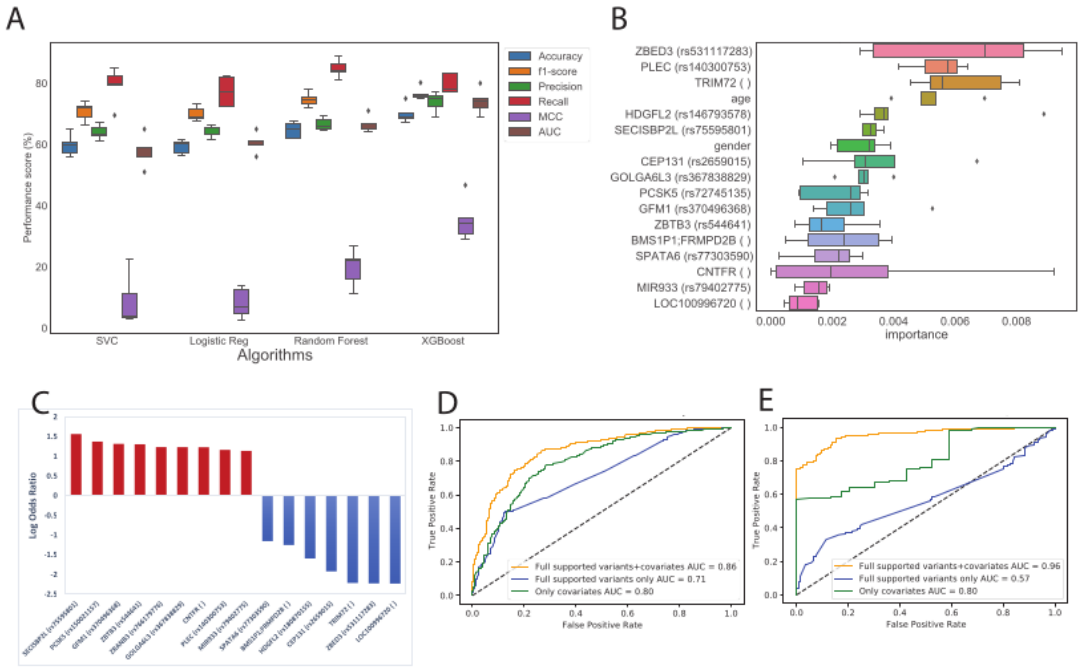

Figure 2

Figure 2

A) distribution of performance metrics of different algorithms during testing on the 5 Folds; B) feature importance distribution for features with non-zero importance across the five folds; C) log-odds ratio of the 16 variants with full support in XGBoost trained models; D) performances of the predictors with 16

variants plus covariates (age and gender; orange), only co-variates (green),all screened variants plus covariates (blue) in the held-out test set; E) performances of the predictors with 16 variants plus covariates (age and gender; orange), only co-variates (green),all screened variants plus covariates (blue) in a follow-up testing set cohort (618 new samples).

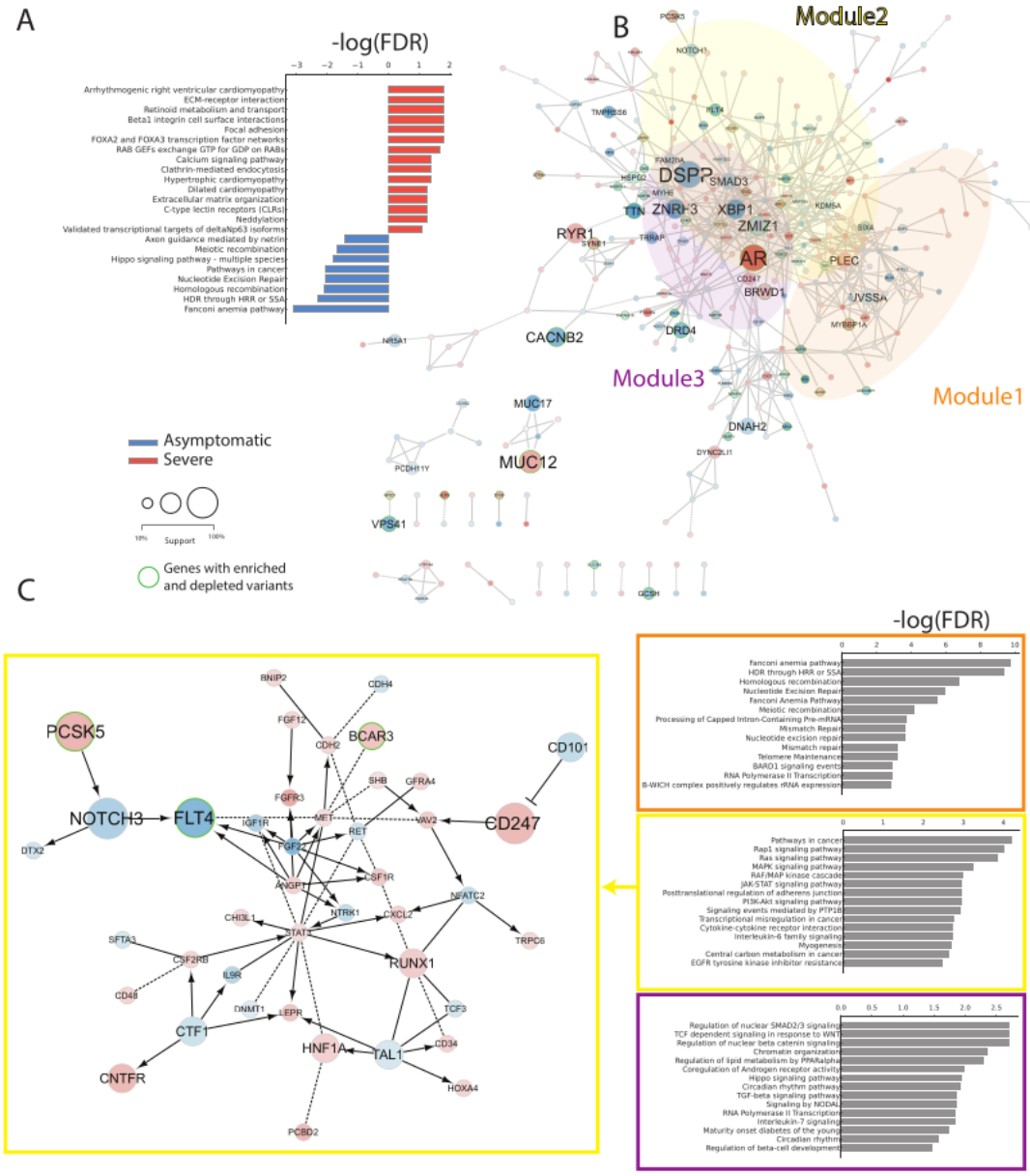

Figure 3

Figure 3

network analysis and pathway enrichment. A) pathways overrepresented among variants with non-zero feature in at least one XGB model and enriched in either severe(red) or asymptomatic (blue); B) Reactome FI network of genes affected by variants with non-zero feature importance from XGBoost. Node diameter is proportional to the number of variants with non-zero coefficients in any decision-tree based models. Node color is instead proportional to the LOR with the highest absolute value among the variants associated to a given gene. The top 3 modules identified within the network are highlighted and corresponding enriched processes displayed as barcharts colored with cluster specific corresponding colors; C) FI network zoomed representation of the 2nd largest cluster

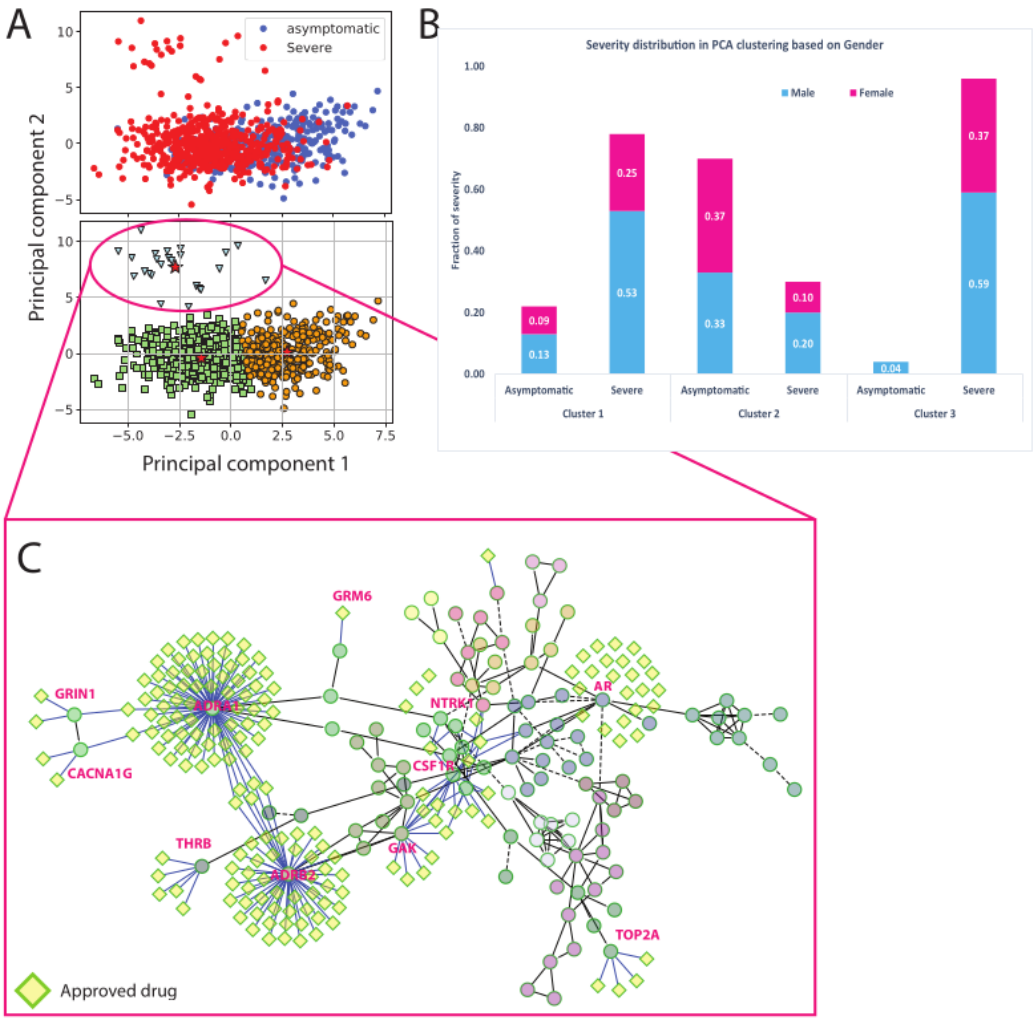

Figure 4

Figure 4

detection of distinct clinical groups via PCA and clustering. A) projection of samples along the 1st and 2nd principal components and coloring based on severity (up) or clusters identified via k-means (bottom); B) gender and clinical group composition of the clusters detected via k-means on the 1st and 2nd PCA components; C) FI network constructed using mutated genes on the cluster of more severe patients and approved drugs available for any of these genes.

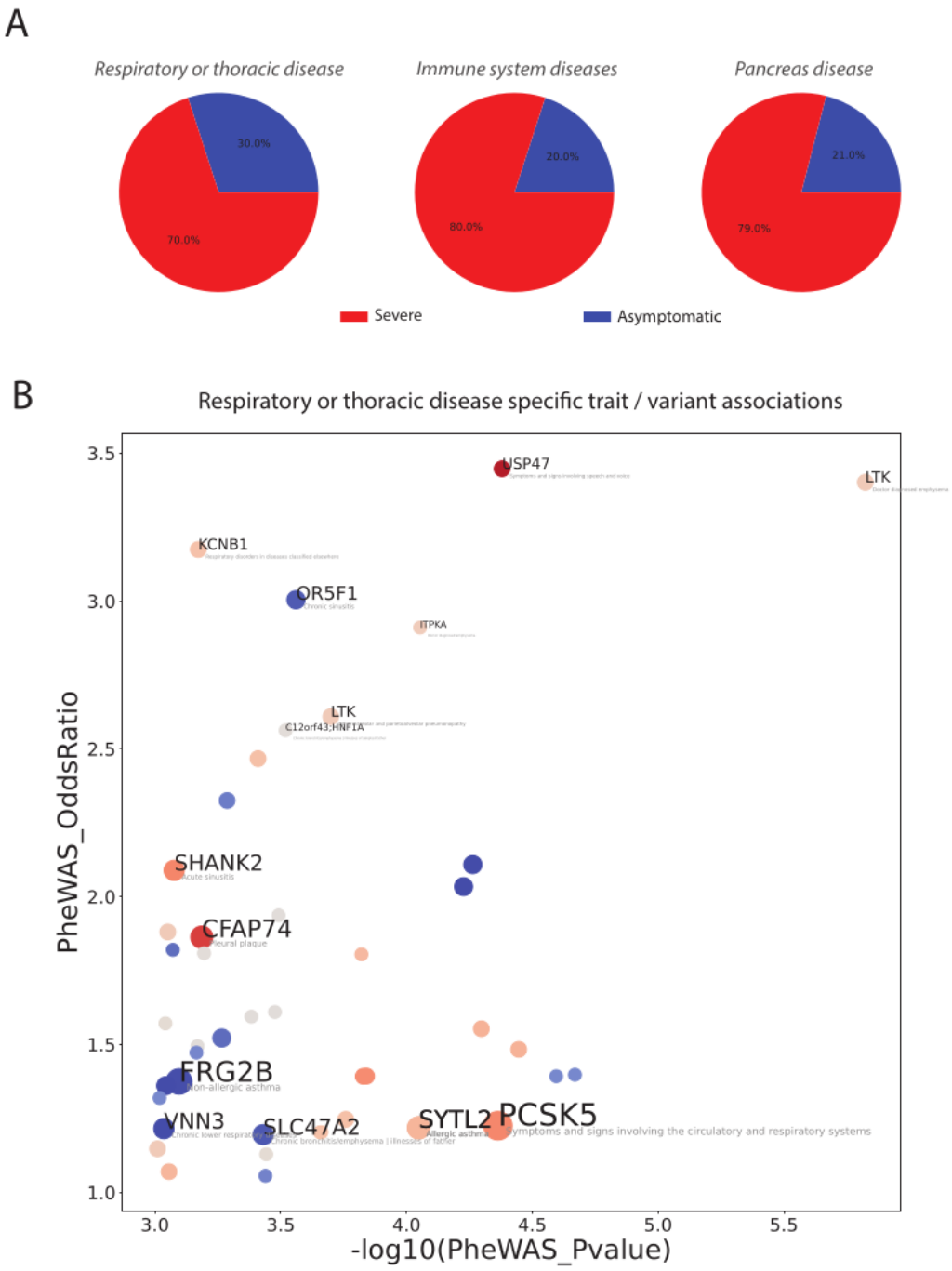

Figure 5

Figure 5

A) phenotype categories displaying the greatest fraction of specific trait associations with variants enriched in severe versus asymptomatic patients; B) scatter plot showing variant-specific traits associated within the “Respiratory or thoracic disease category”. Dot diameter is proportional to the model support for each variant. The color is proportional to the log-odds ratio of the variant in the two

groups of the cohort. Labels are printed only associations with PheWAS Pvalue < 0.001 and PheWAS oddsratio > 2.5 or for variants having non-zero coefficients in at least one XGBoost model.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTables.xlsx
- SupplementaryFigures.pdf