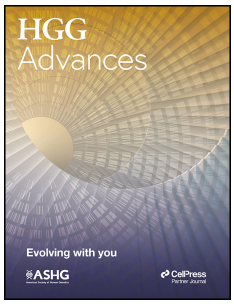
Journal Pre-proof



Stratified analyses refine association between *TLR7* rare variants and severe COVID-19

Jannik Boos, Caspar I. van der Made, Gayatri Ramakrishnan, Eamon Coughlan, Rosanna Asselta, Britt-Sabina Löscher, Luca V.C. Valenti, Rafael de Cid, Luis Bujanda, Antonio Julià, Erola Pairo-Castineira, J. Kenneth Baillie, Sandra May, Berina Zametica, Julia Heggemann, Agustín Albillos, Jesus M. Banales, Jordi Barretina, Natalia Blay, Paolo Bonfanti, Maria Buti, Javier Fernandez, Sara Marsal, Daniele Prati, Luisa Ronzoni, Nicoletta Sacchi, The Spanish/Italian Severe COVID-19 Sequencing group, GenOMICC Investigators, Joachim L. Schultze, Olaf Riess, Andre Franke, Konrad Rawlik, David Ellinghaus, Alexander Hoischen, Axel Schmidt, Kerstin U. Ludwig

PII: S2666-2477(24)00063-0

DOI: https://doi.org/10.1016/j.xhgg.2024.100323

Reference: XHGG 100323

To appear in: *Human Genetics and Genomics Advances*

Received Date: 16 February 2024

Accepted Date: 25 June 2024

Please cite this article as: Boos J, van der Made CI, Ramakrishnan G, Coughlan E, Asselta R, Löscher B-S, Valenti LVC, de Cid R, Bujanda L, Julià A, Pairo-Castineira E, Baillie JK, May S, Zametica B, Heggemann J, Albillos A, Banales JM, Barretina J, Blay N, Bonfanti P, Buti M, Fernandez J, Marsal S, Prati D, Ronzoni L, Sacchi N, The Spanish/Italian Severe COVID-19 Sequencing group, GenOMICC Investigators, Schultze JL, Riess O, Franke A, Rawlik K, Ellinghaus D, Hoischen A, Schmidt A, Ludwig KU, Stratified analyses refine association between *TLR7* rare variants and severe COVID-19, *Human Genetics and Genomics Advances* (2024), doi: https://doi.org/10.1016/j.xhgg.2024.100323.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that,

during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2024

Journal Pre-proof

Stratified analyses refine association between *TLR7* rare variants and severe COVID-19

Jannik Boos[1], Caspar I. van der Made[2], Gayatri Ramakrishnan[3], Eamon Coughlan[4,5], Rosanna Asselta[6,7], Britt-Sabina Löscher[8], Luca V. C. Valenti[9,10], Rafael de Cid[11,12], Luis Bujanda[13,14], Antonio Julià[15], Erola Pairo-Castineira[4,5], J. Kenneth Baillie[4,5], Sandra May[8], Berina Zametica[1], Julia Heggemann[1], Agustín Albillos[14,16], Jesus M. Banales[13,14,17,18], Jordi Barretina[11], Natalia Blay[11,12], Paolo Bonfanti[19], Maria Buti[14], Javier Fernandez[20,21], Sara Marsal[15], Daniele Prati[10], Luisa Ronzoni[10], Nicoletta Sacchi[22], The Spanish/Italian Severe COVID-19 Sequencing group[23], GenOMICC Investigators[24], Joachim L. Schultze[25,26,27], Olaf Riess[28,29], Andre Franke[8], Konrad Rawlik[4], David Ellinghaus[8], Alexander Hoischen[2], Axel Schmidt[1,#], Kerstin U. Ludwig[1,#,*]

1: Institute of Human Genetics, University of Bonn School of Medicine & University Hospital Bonn, Bonn, Germany
2: Department of Human Genetics; Department of Internal Medicine; Radboudumc Research Institute for Medical Innovation; Radboud Center for Infectious Diseases (RCI), Radboud University Medical Center, Nijmegen, The Netherlands
3: Department of Medical Biosciences, Radboud University Medical Center, Nijmegen, The Netherlands
4: Baillie Gifford Pandemic Science Hub, Centre for Inflammation Research, Institute for Regeneration and Repair, University of Edinburgh, Edinburgh, UK
5: Roslin Institute, University of Edinburgh, Edinburgh, UK
6: Department of Biomedical Sciences, Humanitas University, Via Rita Levi Montalcini 4, 20090 Pieve Emanuele, Milan, Italy
7: IRCCS Humanitas Research Hospital - via Manzoni 56, 20089 Rozzano, Milan, Italy
8: Institute of Clinical Molecular Biology, Kiel University and University Medical Center, Kiel, Germany
9: Department of Pathophysiology and Transplantation, Università degli Studi di Milano, Milan, Italy
10: Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy
11: Genomes for Life-GCAT Lab, CORE Program. Germans Trias i Pujol Research Institute (IGTP), 08916 Badalona, Spain
12: Grup de Recerca en Impacte de les Malalties Cròniques i les seves Trajectòries (GRIMTra) (IGTP), Badalona, Spain
13: Department of Liver and Gastrointestinal Diseases, Biodonostia Health Research Institute, Donostia University Hospital, University of the Basque Country (UPV/EHU), San Sebastian, Spain
14: Centre for Biomedical Network Research on Hepatic and Digestive Diseases (CIBEREHD), Instituto de Salud Carlos III, 28029 Madrid, Spain
15: Vall d'Hebron Hospital Research Institute, Barcelona, Spain
16: Department of Gastroenterology, Hospital Universitario Ramón y Cajal, Instituto Ramón y Cajal de Investigación Sanitaria (IRYCIS), University of Alcalá, Madrid, Spain
17: IKERBASQUE, Basque Foundation for Science, Bilbao, Spain
18: Department of Biochemistry and Genetics, School of Sciences, University of Navarra, Pamplona, Spain
19: Division of Infectious Diseases, Università degli Studi di Milano Bicocca, Fondazione San Gerardo dei Tintori, Monza, Italy
20: Hospital Clinic, University of Barcelona, Barcelona, Spain
21: European Foundation for the Study of Chronic Liver Failure (EF CLif), Barcelona, Spain
22: IBMDR, E.O. Ospedali Galliera, Genova, Italy
23: Full list of authors available in Table S9.
24: Full list of authors available in Table S9.
25: Systems Medicine, Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE) e.V., Bonn, Germany
26: Genomics and Immunoregulation, Life & Medical Sciences (LIMES) Institute, University of Bonn, Bonn, Germany
27: PRECISE Platform for Genomics and Epigenomics, Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE) e.V. and University of Bonn, Bonn, Germany.
28: Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen, Germany
29: DFG NGS Competence Center Tübingen (NCCT), University of Tübingen, Tübingen, Germany
[#] These authors contributed equally.
* Corresponding author: Kerstin U. Ludwig, kerstin.ludwig@uni-bonn.de

Journal Pre-proof

Abstract

Despite extensive global research into genetic predisposition for severe COVID-19, knowledge on the role of rare host genetic variants and their relation to other risk factors remains limited. Here, 52 genes with prior etiological evidence were sequenced in 1,772 severe COVID-19 cases and 5,347 population-based controls from Spain/Italy. Rare deleterious *TLR7* variants were present in 2.4% of young (<60 years) cases with no reported clinical risk factors (n=378), compared to 0.24% of controls (odds ratio (OR)=12.3, $p=1.27x10^{-10}$). Incorporation of the results of either functional assays or protein modeling led to a pronounced increase in effect size (OR_{max}=46.5, $p=1.74x10^{-15}$). Association signals for X-chromosomal *TLR7* were also detected in the female-only subgroup, suggesting the existence of additional mechanisms beyond X-linked recessive inheritance in males. Additionally, supporting evidence was generated for a contribution to severe COVID-19 of the previously implicated genes *IFNAR2*, *IFIH1* and *TBK1*. Our results refine the genetic contribution of rare *TLR7* variants to severe COVID-19, and strengthen evidence for the etiological relevance of genes in the interferon signaling pathway.

Keywords: SARS-CoV-2, host genetics, Toll-like receptor 7, targeted sequencing, rare variants, variant collapsing analysis, burden analysis, innate immunity, immune deficiency, infection

Journal Pre-proof

Introduction

The SARS-CoV-2 pandemic has posed major challenges to societies and health care systems around the world. Clinically, SARS-CoV-2 infection results in a broad spectrum of outcomes, ranging from the complete absence of symptoms to severe illness and even death secondary to the associated lung disease (severe COVID-19). Extensive research has been conducted to elucidate the causes of these inter-individual differences, with the aim of informing drug development programs and designing strategies for individual risk prediction in future viral pandemics. This has demonstrated that the observed variability is explained in part by demographic and clinical risk factors. Specifically, increased age, male sex, comorbidities like diabetes, coronary artery disease (CAD), high body weight, and hypertension[1–3], as well as the presence of auto-antibodies[4] have been suggested to be associated with severe COVID-19. In addition, research has shown robust associations between severe COVID-19 and common genetic variants in the host, which are typically characterized by a minor allele frequency (MAF) of >1% and modest effect sizes[5–10].

Monogenic causes have been suggested in individuals with severe COVID-19, as based on the identification of highly penetrant pathogenic variants in *TLR7* [OMIM: 300365], *TBK1* [OMIM: 604834], and *IFNAR1* [OMIM 107450] in individual families[11–14]. To date, only a limited number of studies have performed systematic investigations of the role of rare genetic variants in large severe COVID-19 cohorts[9,15,16]. At the population level, the most compelling evidence for this to date has been reported for rare variants in the X-chromosomal gene *TLR7*[17–22]. The corresponding protein TLR7 (toll-like-receptor 7) is a receptor for single-stranded RNA and is central to SARS-CoV-2 host defense[23]. The suggested pathomechanism of *TLR7* rare variants in males with severe COVID-19 is X-linked recessive loss-of-function[19]. Since *TLR7* escapes X-inactivation[24], this hypothesis does not explain recent findings of rare deleterious *TLR7* variants in females with severe COVID-19[15].

Given prior epidemiological evidence for a contribution of age, sex, and additional clinical risk factors to the risk for severe COVID-19, the aim of the present study was to empower the search for rare variant associations by performing stratified analyses in two ethnically homogeneous cohorts. For this purpose, 52 candidate genes for severe COVID-19, including *TLR7*, were sequenced in 1,772 individuals from Spain and Italy who had been hospitalized for COVID-19 and had required respiratory support, and 5,347 individuals from the general Spanish/Italian populations. Notably, the severe COVID-19 cases were recruited prior to vaccine availability, thus allowing analysis of the virus-naive host reaction to SARS-CoV-2 infection. All individuals had undergone previous array-based genotyping as part of prior genome-wide association studies (GWAS)[25,26]. The candidate gene sequencing approach was based on the cohort's informed consent on targeted follow-up sequencing. Together with available clinical information, sequencing data were then analyzed for single variant associations and gene burden using different stratified approaches, including distinct phenotype definitions and variant pathogenicity levels.

Subjects and Methods

Candidate gene selection

The available informed consent documentation allowed follow up sequencing only, and precluded systematic approaches such as exome sequencing (ES). Therefore, 55 genes were selected in August 2020, based on evidence available at that time. These comprised 14 genes from early GWAS loci[5,25]; five genes from diagnostic ES[11,13]; and 36 genes with functional evidence, which have been implicated previously in viral defense or pathogen immunity (Figure 1a). For each gene, the evidence for selection is presented in Table S1. Three genes (*CCL3*, *CXCL1*, *CFD*) were subsequently excluded from the analysis, since the size of the respective covered region post quality control (QC) was less than 50% of the

originally targeted region. Detailed information on the coverage of these genes, and the number of variant sites per gene, is provided in Table S1.

Study design and phenotype definition

Coding regions were sequenced using single-molecule Molecular Inversion Probes (MIPs)[27], in 9,104 Spanish/Italian individuals from the Severe COVID-19 GWAS cohort[25,26] (see Supplemental Methods). Following post-sequencing QC, which included the use of array-based genotype data for population inference and relatedness filtering (Figure 1b), a total of 7,119 individuals remained for analysis. Data analysis included (i) single variant association analysis and (ii) rare variant collapsing analysis. Both analyses were performed using four case-control definitions (Table 1) which involved one main analysis comprising the entire cohort, and three stratified analyses. The stratified analyses were performed in order to investigate the contribution of rare variants in individuals with otherwise low epidemiological risk ($POP_{lowrisk}$, COV_{hosp} by risk factors), and the potential contribution of rare variants to the level of disease severity (COV_{hosp} by respiratory support). Each of the four analyses was repeated separately for males and females, in view of prior reports of sex-differences in etiological risk[3]. Notably, some COV_{hosp} individuals (66 of 1,772) did not have sufficient information on comorbidities and were therefore excluded from the risk factor based stratifications ($POP_{lowrisk}$, COV_{hosp} by risk factors).

Cohort characteristics

The recruitment procedure, sample collection, and DNA extraction were conducted by the Severe COVID-19 GWAS group (Figure 1b) and are described elsewhere[26]. Approvals were obtained from the relevant ethics committees (listed in Supplemental Methods), and informed consent was obtained. Individuals hospitalized for severe COVID-19 (COV_{hosp}) were collected at several centers in Spain and Italy in 2020 as part of the first outbreaks of the pandemic in Europe. Severe COVID-19 was defined as requiring respiratory support, i.e. the necessity for oxygen supplementation. While other definitions exist, this approach was

chosen to ensure feasibility[26]. Following QC (see next paragraph) the cohort comprised: (i) 1,772 COV_{hosp} individuals (n=1,008 from Italy, n=764 from Spain; Figure 1c, Table 1); and (ii) 5,347 population-based controls (n=1,408 from Italy, n=3,939 from Spain). In total, 38% of all individuals were female. Respiratory support for COV_{hosp} individuals was documented as the maximum support required during hospitalization: oxygen mask only (level 1, lowest), non-invasive ventilation (level 2), invasive ventilation (level 3), or extracorporeal membrane oxygenation (ECMO, level 4, highest). For the majority of the COV_{hosp} individuals (1,706 of 1,772) data were available on comorbid CAD, diabetes, and hypertension (see Figure S1 for further information including subcohort (Italy/Spain) specific distribution of risk factors).

Quality control and data processing

After library preparation and sequencing using MIPs[27] (2x150 bp, paired-end, see Supplemental Methods), data were processed using an MIP-specific pipeline that included several filter and QC steps (Supplemental Methods) and various tools[28–34]. DNA QC, population inference, and relatedness filtering had been performed previously by the Severe COVID-19 GWAS Group[25,26] using their array-based genotype data.

Two patients in the Asano et al.[19] study had phenotypes, age, sex, and rare *TLR7* variants that were identical to those in the present data, suggesting a sample overlap. After recontacting the groups responsible for the recruitment of these two individuals, a total of 82 individuals who may have been common to other research groups were identified. Rare *TLR7* variants of previously reported individuals are labeled accordingly (Figure 5, Table S2).

Single variant analysis

Analysis of the present cohort. An additive non-singleton single variant association test was performed using logistic regression with plink[35] v2.0 and firth correction, as well as age, sex, age^2, age*sex, and the first 10 principal components (PCs) as covariates. The number of PCs was chosen in accordance with Degenhardt et al.[26] and the COVID-HGI exome wide

association study[15]. As the target region spans only about 0.003% of the human genome, the PCs were calculated using the respective array-based genome-wide genotype data (Degenhardt et al.[26]) to maximize the capture of population structure. As case-control ratios and other sample characteristics were substantially different between both populations, logistic regression was performed separately for the Italian and the Spanish cohorts, and the results were then meta-analyzed using METAL[36]. We applied two thresholds for multiple testing: The "strict" threshold was established using the Bonferroni method, which involved correction for the number of analyses (four case-control definitions, three sex-based stratifications) and the number of tested variants (strict, $\alpha=6.7\times10^{-6}$). To take the potential correlation of the different analyses into account, a "lenient" significance threshold was applied, involving correction for the number of tested variants only (lenient, $\alpha=4.1\times10^{-5}$).

Replication cohorts. Whenever COVID-HGI release 7 analysis A2 summary statistics[7] were used as replication cohort, this refers to the leave-one-out-HOSTAGE dataset (which excludes all individuals who were common to the present cohort and the COVID-HGI). For comparison and meta-analysis of the present single variant association results with those of the Regeneron dataset[16], the results of the POP_{all} analysis and the $POP_{lowrisk}$ analysis (without sex stratification) were followed up for all variants with OR>5 and p<0.05 in our cohort. When associations for these variants were reported in the Regeneron Browser (see Web Resources), the respective results were filtered for: (i) the use of Exome data (instead of imputed data); (ii) a phenotype corresponding to that used in the present study ("COVID-19 positive severe vs COVID-19 negative or COVID-19 status unknown" or "COVID-19 positive hospitalized vs COVID-19 negative or COVID-19 status unknown", as defined in Kosmicki et al.[16]); (iii) "European" or "pan-ancestry" ancestry; and (iv) the analysis type "meta-analysis". For each variant, the results of the analysis that included the maximal number of cases were selected.

Gene-based rare variant collapsing analysis

Variant collapsing (or: burden testing) is a widely used approach that is applied to increase statistical power for the testing of rare variants. Here, variants from distinct genetic regions (e.g., in the present study, genes or gene groups) are combined, and testing is performed for these variant groups rather than for single variants.

Definition of variant classes. The present analyses considered two allele frequency groups: MAF<1% and MAF<0.1% (defined as maximal MAF in this cohort or in gnomAD r2.1 non-Finnish European (NFE) exomes). Cohort allele frequencies were calculated using plink v2.0. Deleteriousness classes SYN, M1, M3, M4, and C10+M1 were used. M1, M3, and M4 are similar to those described in Kosmicki et al.[16]. The M1 class is restricted to pLoF variants that are defined as having an Ensembl Variant Effect Predictor (VEP)[33] impact of "HIGH". M3 contains all M1 variants, plus variants with a VEP impact of "moderate" but not missense and missense variants for which five of five prediction algorithms (SIFT, PolyPhen2-HDIV database, PolyPhen2-HVAR database, LRT, MutationTaster) predict deleteriousness. M4 contains all M3 variants plus missense variants for which at least one of the five algorithms predicts a deleterious effect. SYN contains synonymous variants only, and functions as a control class. C10+M1 contains all pLoF (M1) variants and all variants with a CADD v1.6[37] (Combined Annotation Dependent Depletion) score greater than 10, as used by Kousathanas et al.[9].

TLR7-specific variant definitions. For *TLR7,* two additional gene-specific deleteriousness classes were created. The first one comprised biochemically loss-of-function (bLoF) variants, i.e. all variants reported as being LoF on the basis of biochemical tests in previous research[18–20]. Synonymous *TLR7* variants were inspected for potential cryptic splicing effects using spliceAI[38]. The second class (3D-P) comprised variants that were deemed pathogenic or likely pathogenic based on protein structural analyses. Herefore, each of the mutation

sites was analyzed in the context of its structural environment and with regards to changes in protein folding stability. The latter analyses aimed to address whether pathogenicity can be inferred from the extent of mutation-induced changes in the structural integrity of the TLR7 dimer (see Supplemental Methods[39–42]).

Statistical analysis. For the statistical analysis of the collapsed variants, the Cochran–Mantel–Haenszel (CMH) test (plink v1.9 implementation, dominant model) was used, as previously described[43]. While other methods exist, the CMH test was chosen as it was developed for case-control studies with subgroups of different characteristics by performing internal stratification while still generating overall test statistics for the entire cohort[44]. Moreover, the CMH test can handle rare events[45], which is especially useful for rare variant collapsing analysis. The stratification categories used for the CMH-test were subcohort (Italy, Spain) and sex (male, female). Similar to the single-variant association analyses, two thresholds for statistical correction were applied: The “strict” definition was performed according to Bonferroni, and accounted for all performed tests (tested genes, variant categories, case-control definitions, $\alpha=8.7\times10^{-6}$). The “lenient” threshold considered that the case-control definitions and the different variant categories, are correlated, and therefore corrected for the number of tested genes only ($\alpha=9.6\times10^{-4}$). Data from the GenOMICC-study[9] was used for a replication attempt, details for which are provided in the Supplemental Methods.

Results

Single variant analyses identify etiological variant in *TBK1*

Within the 52 genes, 3,218 high-confidence variants were identified across the entire cohort, 95% of which were rare (n=3,059; MAF < 1%). Of these rare variants, 28.6% had no reported frequency in gnomAD r2.1 exomes (n=874, Figure 1d, Figure S2). More specifically, 2,007 singletons (i.e., variants that occur in only one individual) were observed,

including 111 putative loss-of-function (pLoF) variants. These were present in 31 COV_{hosp} individuals, and 77 population-based controls (1.75% vs 1.44%; three individuals carried two variants, respectively). Within the subset of COV_{hosp} individuals with no reported risk factors, eight singleton pLoFs were observed in seven individuals (1.85%), all of which were heterozygous and two of which were found in one individual (Table S3). For these seven individuals, the distribution of age and level of respiratory support did not differ significantly from those of the remaining COV_{hosp} individuals with no reported risk factors (Welch's $p>0.39$).

Next, formal association testing for the 1,211 non-singleton variants was performed using Firth's logistic regression and the covariates age, sex, age^2, age*sex, and 10 principal components (PCs) obtained from prior array-based genotyping (see Methods). This was performed separately for the Spanish and Italian cohorts, and the results were meta-analyzed using inverse variance based meta-analysis (Figure 2, Figure S3). Overall, seven variants had p-values below the strict significance threshold (see Methods). All of these seven variants were associated at genome-wide significance (and with the same direction of effect) in the independent data freeze of the global COVID-19 Host Genetics Initiative (HGI)[7] (release 7, see Methods). Variants associated with nominal significance ($p < 0.05$) and gnomAD r2.1 NFE exomes-AF > 0.01% are reported in Table S4.

Given the limited statistical power for single-variant analyses, candidate variants (defined as high effect size estimates (OR > 5) and nominal significance ($p < 0.05$)) from the POP_{all} and $POP_{lowrisk}$ analyses (non sex-stratified) were followed-up in the Regeneron dataset (see Methods). A total of 62 variants, all of which had an MAF < 0.2% and were absent from the COVID-19 HGI data, met these criteria. Of those, 38 variants were also present in the Regeneron dataset (Table S5). The most significant variant was a missense variant in *TBK1* (p.Arg358His, chr12:64878163:G:A (hg19), CADD=23.3, REVEL=0.259), which showed effect sizes of >20 in both cohorts (Regeneron: OR=24.2, confidence interval = [3.64,

160.47], p=0.00097; present study: OR=30.0 [2.71, 332.6], p=0.0056). In a meta-analysis of both cohorts, this variant showed strong association with severe COVID-19 ($p=1.67x10^{-05}$, OR=26.3 [5.93, 116.2]).

Gene-based rare variant collapsing analysis confirms *TLR7* association

To increase statistical power, gene-based collapsing analyses were performed. For this purpose, variants were assigned to: (i) two allele frequency groups (MAF<0.1% and MAF<1%); and (ii) five classes of deleteriousness (M1, M3, M4, C10+M1, SYN; see Methods). Variant counts per class are provided in Figure S2. For each combination of MAF, deleteriousness and gene, statistical association analyses were performed using the CMH test. The results are reported in Figure 3 for MAF<0.1% and in Figure S4 for both MAF<1% and sex-stratified analyses, respectively. At strict threshold definition (Methods), significant associations were obtained for *TLR7* in: (i) the $POP_{lowrisk}$ analysis overall (C10+M1, MAF<0.1%; carriers: 9/378 cases vs. 13/5347 controls; $p=1.27x10^{-10}$, OR=12.3 [4.7, 32.2]; Figure 3); and (ii) the female-only subgroup (C10+M1, MAF<0.1%; 4/126 vs. 5/2102; $p=1.75x10^{-09}$, OR=24.8 [5.9, 105.2]; Figure S4). Suggestive evidence (at lenient threshold, see Methods) was obtained for two additional genes: (i) *IFNAR2* [OMIM: 602376] (POP_{all}, C10+M1, MAF<1%; 60/1772 vs. 73/5347; $p=2.61x10^{-04}$, OR=1.9 [1.3, 2.7]; Figure S4); and (ii) *IFIH1* [OMIM: 606951] (COV_{hosp} by respiratory support, C10+M1, MAF<1%; 54/478 vs. 36/661; $p=3.60x10^{-04}$, OR=2.2 [1.4, 3.4]; Figure S4). All associations with nominal significance (p<0.05) are listed in Table S6.

To investigate whether genes with related functions were enriched for rare variants, eight gene sets were defined (Table S7), and a collapsing analysis based on each gene set was conducted. No significant results were obtained after strict correction for multiple testing (Figure S5). Nevertheless, the most significant associations were observed for the set of immunodeficiency genes (n=15), and this remained nominally significant even after the exclusion of *TLR7*.

Identification of a low-frequency *TLR7* risk variant in the Spanish population

In view of both the highly significant results presented above and robust prior evidence for the involvement of *TLR7* in severe COVID-19[11,12,18–21], more detailed investigations were performed to characterize the rare *TLR7* variant associations in the present cohort. Overall, the identified *TLR7* variants comprised 26 missense, one 3'-UTR, and 16 synonymous (max. spliceAI score: 0.02) variants, but no pLoF variants (see Table S2). Two COV_{hosp} individuals (one male case, one female case; none of the population-based controls) carried two distinct rare variants in *TLR7* respectively. The male individual (p.M854I, p.L988S) was previously reported in an independent study by Asano et al.[19] (see Methods). In the female individual, biallelic occurrence of the two deleterious variants (p.A448V, p.R920K) could cause X-linked recessive disease. While no direct assessment of compound heterozygosity based on MIP sequencing data was possible, *in silico* haplotype assessment using the variant co-occurrence tool of gnomAD v2.1.1 (see Web Resources)[46] suggested that the two variants map to different haplotypes.

The analyses also identified a missense variant exclusive to the Spanish subcohort (rs202129610, p.D332G). This was present in two population-based controls (MAF=0.038%, one female, one male), and three COV_{hosp} individuals (MAF=0.33%, one female, two males), The frequency further increased in COV_{hosp} individuals with no reported risk factors (MAF=1.0%). The variant was nominally significant in the single variant logistic regression analysis ($POP_{lowrisk}$, OR=5.77 [1.49, 22.3], p=0.011), but was absent from the Regeneron dataset and the *in silico* pathogenicity prediction of this variant was ambiguous (CADD=18.45, REVEL=0.078). However, a previous study reported that this variant was hypomorphic, as based on *in vitro* experiments (7% NfκB activity[19]). This variant is absent from European individuals in gnomAD v3.1.2, and has only been reported to date in Latino/Admixed Americans (population-specific MAF of 0.019%).

Incorporation of functional and protein data increases *TLR7* rare variant effect sizes

Seventeen of the 26 *TLR7* missense variants have previously been analyzed *in vitro*. In these experiments, seven variants were reported to decrease or even abolish the function of TLR7[18–20]. These seven variants were combined to a new deleteriousness class (bLoF, biochemically loss-of-function, as proposed in Matuozzo et al.[21]) for the rare variant collapsing analysis. The resulting OR ($POP_{lowrisk}$, bLoF, MAF<0.1%; 4/378 vs. 3/5347; $p=1.73x10^{-10}$, OR=34.6 [6.8,177.2]; Figure 4a) was substantially higher than effect sizes based on *in silico* prediction alone ($POP_{lowrisk}$, C10+M1, MAF<0.1%; OR=12.3; see above).

To create a structure-based variant class, protein structure data for TLR7 were used for 3D modeling and protein energy calculation (Methods, Supplemental Methods). Based on this approach, eight of the 26 rare missense variants were classified as either damaging (n=4) or probably damaging (n=4) to the protein structure, and were aggregated into a new variant class (3D-P). Statistical analysis of this 3D-P class yielded even higher ORs ($POP_{lowrisk}$, 3D-P, MAF<0.1%; 7/378 vs. 4/5347; $p=1,74x10^{-15}$, OR=46.5 [10.9, 198.7]) than the aforementioned variant classifications (see Figure 4a and Figure S6). In alignment with prior studies that identified *TLR7* associations in younger individuals[11,12,18–21], the analysis was repeated by defining cases as individuals with severe COVID-19 aged < 60 years, with no consideration of other risk factors, and comparing these individuals to all population-controls. Using the 3D-P *TLR7* (MAF<0.1%) class, the proportion of carriers increased across the following three subgroups: all COV_{hosp} individuals (0.45%); younger COV_{hosp} individuals (age < 60 years, 1.25%); COV_{hosp} individuals with no reported risk factors (1.85%; Figure 4b).

Investigation of domain- and sex-specific variant effects in *TLR7*

To date, X-linked TLR7 deficiency, as mediated by rare deleterious variants, has mainly been reported in males[11,12,18–20], and a classical X-linked recessive mode of inheritance has been suggested[11,12,18,19]. However, two recent association studies also reported an enrichment of rare variants in females[15,21]. Given the present finding of an enrichment of rare

heterozygous *TLR7* variants in females, and previous observations of *TLR7* escaping X-inactivation in immune cells[24], analyses were performed to explore other potential mutational mechanisms. First, the distribution of deleterious rare variants across the TLR7 protein was studied in females with no reported risk factors (i.e., $POP_{lowrisk}$; C10+M1, MAF<0.1%). In female cases, an overrepresentation of these variants was observed in the leucine-rich-repeat (LRR) domain (see Figure 5a). Since the LRR domain is involved in the dimerization of TLR7 monomers, which is essential for the activation of downstream signaling pathways[47], we hypothesized that missense variants located in this domain could potentially confer a dominant-negative effect by affecting protein dimerization. We approached this by using the TLR7 protein structure, and observed that four non-synonymous variants (Q138R, H298R, H630Y, I759V; all singleton, all missense) in the entire cohort were located within 5 Ångström of the dimerization interface (I5AN; hashed residue labels in Figure 5). Two of these I5AN variants (Q138R, H630Y) were present in female COV_{hosp} individuals with no reported risk factors, and were among the 3D-P variants (indicating a damaging structural effect, see above). No I5AN variant was observed in controls ($POP_{lowrisk}$ females, I5AN, MAF<1%; 2/126 vs. 0/2101; $p=2.1x10^{-6}$; Figure 5). The two other variants (H298R, I759V) were observed in male controls ($POP_{lowrisk}$ males, I5AN, MAF<1%; 0/252 vs. 2/3245; p=0.65).

To replicate the domain- and sex-specific *TLR7* findings, analyses were performed in the cohort of the GenOMICC study, which has generated one of the largest collections of genome sequencing (GS) data from individuals with severe COVID-19 to date[9]. Overall, only very few numbers of *TLR7* variants were observed in females, and no I5AN variant was observed in either female cases or controls. Detailed results are shown in Table S8 and methodical information is presented in the Supplemental Methods.

Discussion

The present study investigated the contribution of rare genetic variants within 52 candidate genes to the etiology of severe COVID-19 and their relation to clinical risk factors, via the performance of joint and stratified analyses in two large, ethnically homogeneous cohorts recruited in the pre-vaccine era of the SARS-CoV-2 pandemic. The present findings reinforce prior genetic evidence for an etiological role of the X-chromosomal gene *TLR7* in severe COVID-19 through the identification of a robust enrichment of deleterious rare variants. Notably, this enrichment was particularly pronounced in young individuals with severe COVID-19 with no reported demographic or clinical risk factors, and was also present in the female-only subgroup. Together with results from protein structural modeling, this suggests the existence of more complex pathomechanisms of *TLR7* variants, beyond X-linked recessive loss-of-function. The analyses also generated statistical evidence that rare variants in three genes of the interferon signaling pathway, specifically *IFNAR2*, *IFIH1,* and *TBK1*, contribute to severe COVID-19, though these findings require further follow-up.

TLR7 is a cytosolic receptor that recognizes single-stranded RNA, and is a central component of the interferon signaling pathway during SARS-CoV-2 host defense[23]. Multiple lines of evidence suggest that deleterious variants within *TLR7* play a causal role in severe COVID-19[11,12,18–21], and this eventually resulted in recognition of TLR7 deficiency as an inborn error of immunity[48] [OMIM 301051]. Research suggests that TLR7 deficiency is more frequent in younger (<60 years) severe COVID-19 cases[21], which is consistent with the hypothesis that the contribution of host genetic factors is larger in young individuals[49], as has been demonstrated for other risk loci for severe COVID-19, e.g., at the key GWAS locus 3p21.31[50]. To refine the subgroup in which severe COVID-19 secondary to TLR7 deficiency is prevalent, the present analyses extended the list of non-genetic risk factors beyond age by including available data on diabetes, hypertension, and CAD. The largest effect size for the

association of rare deleterious *TLR7* variants with severe COVID-19 was observed in young individuals with none of the aforementioned risk factors. Specifically, in these cases, an approximately 10-fold increase in the proportion of individuals carrying variants that were predicted to be deleterious was observed (2.4% vs. 0.24% in population based controls, C10+M1, MAF<0.1%). Variant classification via 3D protein structural analysis (3D-P, MAF<0.1%) further refined this overrepresentation to 1.85% in young individuals with severe COVID-19 and none of the listed risk factors, compared to 0.07% in population-based controls.

In the female-only subgroup, the present analyses identified a strong enrichment of rare *TLR7* variants that were predicted to be damaging. While such an enrichment has been observed in previous independent cohorts[15,21], the underlying mechanisms were not explored. The proposed X-linked recessive model[19] suggests that TLR7 deficiency would be restricted to females with biallelic deleterious mutations. While we identified one female with presumed compound heterozygosity, this individual was not among the cases of the $POP_{lowrisk}$ analysis and did not contribute to the observed burden. We therefore suggest the existence of an additional pathomechanism in heterozygous females, which may be dominant-negative in nature. We hypothesized that an affected TLR7 monomer would interfere with dimerization, thereby reducing TLR7 function by >50%. In support of this, an overrepresentation of *TLR7* missense variants that surrounded the dimerization interface in 3D space was observed in female cases. This observation adds to accumulating evidence for an allelic series underlying TLR7 dosage and its relevance to human immune disorders. The most recent support for this was provided by reports of hypermorphic or gain-of-function mutations in TLR7, which underlie monogenic forms of systemic lupus erythematosus[51] [OMIM 301080]. However, we were unable to obtain additional confirmation from the GenOMICC cohort due to power limitations, such as the very low number of variant observations and the differing cohort characteristics, including recruitment criteria. Future

functional *in vitro* investigation of the pathogenic variants that were found in the present female cases are required to confirm our hypothesis.

The present analyses also identified a missense *TLR7* variant (rs202129610, p.D332G) that was specific to the Spanish subcohort. This variant, which has *in vitro* evidence for deleteriousness[19], was observed in 3 of 764 severe COVID-19 cases from Spain (MAF=0.33%), including 2 out of 147 young hospitalized individuals with no additional risk factors (MAF=1.0%). This is substantially higher than the allele frequency observed in the present Spanish controls (MAF=0.038%), as well as estimates from the Latin American population groups from the gnomAD data v3.1.2 (0.019%).

Besides the results for *TLR7*, the present analyses generated several other interesting findings that require replication in larger cohorts. Specifically, associations with severe COVID-19 were found for *IFNAR2* and *IFIH1* in the rare variant collapsing analysis and for a rare missense *TBK1* variant in the single variant analysis. All of the three genes are involved in the interferon signaling pathway[23], and prior evidence for involvement in severe COVID-19 has been presented[13,52,53]. The observed rare *TBK1* missense variant (p.Arg358His) was found in two of 378 young cases with no reported risk factors and only one of 5,347 controls. Although statistical evidence for this variant was not robust to multiple testing in our study alone, its independent replication in the Regeneron dataset adds to the prior finding of a rare deleterious *TBK1* variant in a child with severe COVID-19[13]. Furthermore, our observation of an enrichment of rare variants in the broader group of immunodeficiency genes, even after the exclusion of *TLR7,* suggests that this set of genes is likely to harbor a substantial proportion of the rare variant risk for severe COVID-19.

While our results contribute to ongoing work into the role of rare variants within the overall host genetic architecture of severe COVID-19, the present study had some inherent limitations. First, the candidate gene approach, which was selected due to a lack of informed

consent for more systematic ES/GS analyses, limited the number of analyzed genes to 52. This prevented identification of additional risk genes, and also poses challenges regarding population substructure which might cause confounding in rare variant studies[54]. To address the latter, we took advantage of the availability of prior array-based genotypes[25,26], which decreased the risk of false-positive findings due to population stratification. Second, gene selection was performed in August 2020, and thus subsequently reported risk genes were not examined, e.g., those located at loci that have been reported in recent global GWAS[7,10]. Third, comorbidity data were limited, and did not include the now well established risk factor increased weight - usually measured as body-mass-index (BMI) - which is one of the strongest clinical predictors of severe COVID-19[3]. However, CAD, diabetes, and hypertension are all correlated with BMI, which suggests that the present analyses captured this effect at least in part. Of note, following initial evidence on hypertension being an independent risk factor for severe COVID-19[2], subsequent studies have reported ambiguous results[55]. Given that individual array-based genotypes are available for the individuals included in the present study, future refinement analyses might include the evaluation of genetically-mediated obesity via the integration of polygenic risk scores. Finally, in the present analysis, the selection of variants with a deleterious effect on protein function was mainly based on computational prediction tools, since (with the exception of some variants within *TLR7*) experimental data on genetic variants are limited. Particularly for missense variants, computational prediction tools are imperfect, and misclassification probably decreased the power of the gene-based collapsing analyses. However, a tailored, molecular modeling approach for missense variants within *TLR7* was used in order to fine-tune the statistical analyses and led to increased effect size estimates. In the future, new approaches, such as novel computational prediction tools that build more strongly on protein structural information[56–58], and data from deep mutational scanning experiments, could improve statistical power, and enhance the information content of the present data.

Journal Pre-proof

Despite the residual open questions, our stratified analysis approach refined the association between rare deleterious *TLR7* variants and severe COVID-19. We suggest a candidate pathomechanism in females, which was identified on the basis of the integration of cohort-level sequencing data and information on protein structure.

Supplemental Information

Supplemental Information includes seven figures, nine tables, Supplemental Methods and Author contributions.

Data and Code Availability

Individual-level data, including raw sequencing data and genotypes, are unavailable for sharing due to consent restrictions. Single variant summary statistics (MAF>0.01%) and the results of the burden analyses are made available at Zenodo (doi: 10.5281/zenodo.11148109). Code used for the analyses in the study is openly available and referenced throughout the paper.

Declaration of Interests

K. U. L. is a co-founder of LAMPseq Diagnostics GmbH. The authors have no other competing interests to declare.

Acknowledgements

We thank Julia Fazaal, Anna Carreras, Alessio Aghemo, Antonio Voza, and Maurizio Cecconi (for laboratory and clinical support), Beatriz Cortés (for data transfer support), and Alberto Mantovani and Stefano Duga (for scientific input). We thank the staff of the Basque Biobank in Spain, members of the COVICAT study group, the staff of GCAT|Genomes for

Journal Pre-proof

Life and the Baillie Gifford / Baillie Gifford Science Pandemic Hub (University of Edinburgh). This research received support from the Solve-RD project (to A.H.; funded from the European Union's Horizon 2020 research and innovation programme (No. 779257)), the BONFOR program of the Medical Faculty, University of Bonn (Gerok stipend to A.S., O-149.0134), the Fondazione IRCCS Ca' Granda "FoGS 2021" genomic study (to L.V.C.V.) and the Banca Intesa San Paolo. G.R. was supported by the Europees Fonds voor Regionale Ontwikkeling (to R.A., EFRO, R0005582). J.K.B. gratefully acknowledges funding support from a Wellcome Trust Senior Research Fellowship (223164/Z/21/Z), UKRI grants MC_PC_20004, MC_PC_19025, MC_PC_1905, MRNO2995X/1, and MC_PC_20029, Sepsis Research (Fiona Elizabeth Agnew Trust), and a BBSRC Institute Strategic Programme Grant to the Roslin Institute (BB/P013732/1, BB/P013759/1). The study was partially funded by the Cariplo Foundation in Milan. It received support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through 286/2020B01 (428994620), LU1944-3/1, and infrastructure support from the DFG Clusters of Excellence 2167 "PrecisionMedicine in Chronic Inflammation (PMI)" (EXC 2167-390884018) and 2151 "ImmunoSensation" (390873048). Sequencing was performed at the West German Genome Center (WGGC; INST 216/981-1) and the NGS Core Facility Bonn.

Journal Pre-proof

Web Resources

https://rgc-covid19.regeneron.com/results

https://gnomad.broadinstitute.org/variant-cooccurrence?dataset=gnomad_r2_1&variant=X-12906386-G-A&variant=X-12904970-C-T

https://www.omim.org/

References

1. Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., Xiang, J., Wang, Y., Song, B., Gu, X., et al. (2020). Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. Lancet Lond. Engl. *395*, 1054–1062. https://doi.org/10.1016/S0140-6736(20)30566-3.

2. Yang, J., Zheng, Y., Gou, X., Pu, K., Chen, Z., Guo, Q., Ji, R., Wang, H., Wang, Y., and Zhou, Y. (2020). Prevalence of comorbidities and its effects in patients infected with SARS-CoV-2: a systematic review and meta-analysis. Int. J. Infect. Dis. IJID Off. Publ. Int. Soc. Infect. Dis. *94*, 91–95. https://doi.org/10.1016/j.ijid.2020.03.017.

3. Williamson, E.J., Walker, A.J., Bhaskaran, K., Bacon, S., Bates, C., Morton, C.E., Curtis, H.J., Mehrkar, A., Evans, D., Inglesby, P., et al. (2020). OpenSAFELY: factors associated with COVID-19 death in 17 million patients. Nature *584*, 430–436. https://doi.org/10.1038/s41586-020-2521-4.

4. Bastard, P., Gervais, A., Le Voyer, T., Rosain, J., Philippot, Q., Manry, J., Michailidis, E., Hoffmann, H.-H., Eto, S., Garcia-Prat, M., et al. (2021). Autoantibodies neutralizing type I IFNs are present in ~4% of uninfected individuals over 70 years old and account for ~20% of COVID-19 deaths. Sci. Immunol. *6*, eabl4340. https://doi.org/10.1126/sciimmunol.abl4340.

5. COVID-19 Host Genetics Initiative (2021). Mapping the human genetic architecture of COVID-19. Nature *600*, 472–477. https://doi.org/10.1038/s41586-021-03767-x.

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

7. Kanai, M., Andrews, S.J., Cordioli, M., Stevens, C., Neale, B.M., Daly, M., Ganna, A., Pathak, G.A., Iwasaki, A., Karjalainen, J., et al. (2023). A second update on mapping the human genetic architecture of COVID-19. Nature *621*, E7–E26. https://doi.org/10.1038/s41586-023-06355-3.

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

9. Kousathanas, A., Pairo-Castineira, E., Rawlik, K., Stuckey, A., Odhams, C.A., Walker, S., Russell, C.D., Malinauskas, T., Wu, Y., Millar, J., et al. (2022). Whole-genome sequencing reveals host factors underlying critical COVID-19. Nature *607*, 97–103. https://doi.org/10.1038/s41586-022-04576-6.

10. Pairo-Castineira, E., Rawlik, K., Bretherick, A.D., Qi, T., Wu, Y., Nassiri, I., McConkey, G.A., Zechner, M., Klaric, L., Griffiths, F., et al. (2023). GWAS and meta-analysis identifies 49 genetic variants underlying critical COVID-19. Nature *617*, 764–768. https://doi.org/10.1038/s41586-023-06034-3.

11. 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., et al. (2020). Presence of Genetic Variants Among Young Men With Severe COVID-19. JAMA *324*, 663–673. https://doi.org/10.1001/jama.2020.13719.

12. 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., et al. (2021). Genetic Screening for TLR7 Variants in Young and Previously Healthy Men With Severe COVID-19. Front. Immunol. *12*, 719115. https://doi.org/10.3389/fimmu.2021.719115.

13. Schmidt, A., Peters, S., Knaus, A., Sabir, H., Hamsen, F., Maj, C., Fazaal, J., Sivalingam, S., Savchenko, O., Mantri, A., et al. (2021). TBK1 and TNFRSF13B

Journal Pre-proof

mutations and an autoinflammatory disease in a child with lethal COVID-19. NPJ Genomic Med. *6*, 55. https://doi.org/10.1038/s41525-021-00220-w.

14. Abolhassani, H., Landegren, N., Bastard, P., Materna, M., Modaresi, M., Du, L., Aranda-Guillén, M., Sardh, F., Zuo, F., Zhang, P., et al. (2022). Inherited IFNAR1 Deficiency in a Child with Both Critical COVID-19 Pneumonia and Multisystem Inflammatory Syndrome. J. Clin. Immunol. *42*, 471–483. https://doi.org/10.1007/s10875-022-01215-7.

15. Butler-Laporte, G., Povysil, G., Kosmicki, J.A., Cirulli, E.T., Drivas, T., Furini, S., Saad, C., Schmidt, A., Olszewski, P., Korotko, U., et al. (2022). Exome-wide association study to identify rare variants influencing COVID-19 outcomes: Results from the Host Genetics Initiative. PLoS Genet. *18*, e1010367. https://doi.org/10.1371/journal.pgen.1010367.

16. Kosmicki, J.A., Horowitz, J.E., Banerjee, N., Lanche, R., Marcketta, A., Maxwell, E., Bai, X., Sun, D., Backman, J.D., Sharma, D., et al. (2021). Pan-ancestry exome-wide association analyses of COVID-19 outcomes in 586,157 individuals. Am. J. Hum. Genet. *108*, 1350–1355. https://doi.org/10.1016/j.ajhg.2021.05.017.

17. Zhang, Q., Bastard, P., Liu, Z., Le Pen, J., Moncada-Velez, M., Chen, J., Ogishi, M., Sabli, I.K.D., Hodeib, S., Korol, C., et al. (2020). Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. Science *370*, eabd4570. https://doi.org/10.1126/science.abd4570.

18. Fallerini, C., Daga, S., Mantovani, S., Benetti, E., Picchiotti, N., Francisci, D., Paciosi, F., Schiaroli, E., Baldassarri, M., Fava, F., et al. (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*, e67569. https://doi.org/10.7554/eLife.67569.

19. Asano, T., Boisson, B., Onodi, F., Matuozzo, D., Moncada-Velez, M., Maglorius Renkilaraj, M.R.L., Zhang, P., Meertens, L., Bolze, A., Materna, M., et al. (2021). X-linked recessive TLR7 deficiency in ~1% of men under 60 years old with life-threatening COVID-19. Sci. Immunol. *6*, eabl4348. https://doi.org/10.1126/sciimmunol.abl4348.

20. Mantovani, S., Daga, S., Fallerini, C., Baldassarri, M., Benetti, E., Picchiotti, N., Fava, F., Gallì, A., Zibellini, S., Bruttini, M., et al. (2022). Rare variants in Toll-like receptor 7 results in functional impairment and downregulation of cytokine-mediated signaling in COVID-19 patients. Genes Immun. *23*, 51–56. https://doi.org/10.1038/s41435-021-00157-1.

21. Matuozzo, D., Talouarn, E., Marchal, A., Zhang, P., Manry, J., Seeleuthner, Y., Zhang, Y., Bolze, A., Chaldebas, M., Milisavljevic, B., et al. (2023). Rare predicted loss-of-function variants of type I IFN immunity genes are associated with life-threatening COVID-19. Genome Med. *15*, 22. https://doi.org/10.1186/s13073-023-01173-8.

22. Zhang, Q., Matuozzo, D., Le Pen, J., Lee, D., Moens, L., Asano, T., Bohlen, J., Liu, Z., Moncada-Velez, M., Kendir-Demirkol, Y., et al. (2022). Recessive inborn errors of type I IFN immunity in children with COVID-19 pneumonia. J. Exp. Med. *219*, e20220131. https://doi.org/10.1084/jem.20220131.

23. van der Made, C.I., Netea, M.G., van der Veerdonk, F.L., and Hoischen, A. (2022). Clinical implications of host genetic variation and susceptibility to severe or critical COVID-19. Genome Med. *14*, 96. https://doi.org/10.1186/s13073-022-01100-3.

24. Souyris, M., Cenac, C., Azar, P., Daviaud, D., Canivet, A., Grunenwald, S., Pienkowski, C., Chaumeil, J., Mejía, J.E., and Guéry, J.-C. (2018). TLR7 escapes X chromosome

inactivation in immune cells. Sci. Immunol. *3*, eaap8855. https://doi.org/10.1126/sciimmunol.aap8855.

25. Severe Covid-19 GWAS Group, Ellinghaus, D., Degenhardt, F., Bujanda, L., Buti, M., Albillos, A., Invernizzi, P., Fernández, J., Prati, D., Baselli, G., et al. (2020). Genomewide Association Study of Severe Covid-19 with Respiratory Failure. N. Engl. J. Med. *383*, 1522–1534. https://doi.org/10.1056/NEJMoa2020283.

26. Degenhardt, F., Ellinghaus, D., Juzenas, S., Lerga-Jaso, J., Wendorff, M., Maya-Miles, D., Uellendahl-Werth, F., ElAbd, H., Rühlemann, M.C., Arora, J., et al. (2022). Detailed stratified GWAS analysis for severe COVID-19 in four European populations. Hum. Mol. Genet. *31*, 3945–3966. https://doi.org/10.1093/hmg/ddac158.

27. Hiatt, J.B., Pritchard, C.C., Salipante, S.J., O'Roak, B.J., and Shendure, J. (2013). Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. Genome Res. *23*, 843–854. https://doi.org/10.1101/gr.147686.112.

28. Boyle, E.A., O'Roak, B.J., Martin, B.K., Kumar, A., and Shendure, J. (2014). MIPgen: optimized modeling and design of molecular inversion probes for targeted resequencing. Bioinforma. Oxf. Engl. *30*, 2670–2672. https://doi.org/10.1093/bioinformatics/btu353.

29. Davis, M.P.A., van Dongen, S., Abreu-Goodger, C., Bartonicek, N., and Enright, A.J. (2013). Kraken: A set of tools for quality control and analysis of high-throughput sequence data. Methods San Diego Calif *63*, 41–49. https://doi.org/10.1016/j.ymeth.2013.06.027.

30. Vasimuddin, Md., Misra, S., Li, H., and Aluru, S. (2019). Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. In 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS), pp. 314–324. https://doi.org/10.1109/IPDPS.2019.00041.

31. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008. https://doi.org/10.1093/gigascience/giab008.

32. Smith, T., Heger, A., and Sudbery, I. (2017). UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res. *27*, 491–499. https://doi.org/10.1101/gr.209601.116.

33. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*, 122. https://doi.org/10.1186/s13059-016-0974-4.

34. Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J.A., Ziyatdinov, A., Benner, C., O'Dushlaine, C., Barber, M., Boutkov, B., et al. (2021). Computationally efficient whole-genome regression for quantitative and binary traits. Nat. Genet. *53*, 1097–1103. https://doi.org/10.1038/s41588-021-00870-7.

35. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience *4*, 7. https://doi.org/10.1186/s13742-015-0047-8.

36. Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. Bioinforma. Oxf. Engl. *26*, 2190–2191. https://doi.org/10.1093/bioinformatics/btq340.

Journal Pre-proof

37. Rentzsch, P., Schubach, M., Shendure, J., and Kircher, M. (2021). CADD-Splice-improving genome-wide variant effect prediction using deep learning-derived splice scores. Genome Med. *13*, 31. https://doi.org/10.1186/s13073-021-00835-9.

38. Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J.F., Darbandi, S.F., Knowles, D., Li, Y.I., Kosmicki, J.A., Arbelaez, J., Cui, W., Schwartz, G.B., et al. (2019). Predicting Splicing from Primary Sequence with Deep Learning. Cell *176*, 535-548.e24. https://doi.org/10.1016/j.cell.2018.12.015.

39. Buß, O., Rudat, J., and Ochsenreither, K. (2018). FoldX as Protein Engineering Tool: Better Than Random Based Approaches? Comput. Struct. Biotechnol. J. *16*, 25–33. https://doi.org/10.1016/j.csbj.2018.01.002.

40. Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005). The FoldX web server: an online force field. Nucleic Acids Res. *33*, W382-388. https://doi.org/10.1093/nar/gki387.

41. Parra, R.G., Schafer, N.P., Radusky, L.G., Tsai, M.-Y., Guzovsky, A.B., Wolynes, P.G., and Ferreiro, D.U. (2016). Protein Frustratometer 2: a tool to localize energetic frustration in protein molecules, now with electrostatics. Nucleic Acids Res. *44*, W356-360. https://doi.org/10.1093/nar/gkw304.

42. Tsai, M.-Y., Zheng, W., Balamurugan, D., Schafer, N.P., Kim, B.L., Cheung, M.S., and Wolynes, P.G. (2016). Electrostatics, structure prediction, and the energy landscapes for protein folding and binding. Protein Sci. Publ. Protein Soc. *25*, 255–269. https://doi.org/10.1002/pro.2751.

43. Cirulli, E.T., Lasseigne, B.N., Petrovski, S., Sapp, P.C., Dion, P.A., Leblond, C.S., Couthouis, J., Lu, Y.-F., Wang, Q., Krueger, B.J., et al. (2015). Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. Science *347*, 1436–1441. https://doi.org/10.1126/science.aaa3650.

44. Mantel, N., and Haenszel, W. (1959). Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease. JNCI J. Natl. Cancer Inst. *22*, 719–748. https://doi.org/10.1093/jnci/22.4.719.

45. Efthimiou, O. (2018). Practical guide to the meta-analysis of rare events. Evid. Based Ment. Health *21*, 72–76. https://doi.org/10.1136/eb-2018-102911.

46. Guo, M.H., Francioli, L.C., Stenton, S.L., Goodrich, J.K., Watts, N.A., Singer-Berk, M., Groopman, E., Darnowsky, P.W., Solomonson, M., Baxter, S., et al. (2023). Inferring compound heterozygosity from large-scale exome sequencing data. Nat. Genet. https://doi.org/10.1038/s41588-023-01608-3.

47. Zhang, Z., Ohto, U., Shibata, T., Krayukhina, E., Taoka, M., Yamauchi, Y., Tanji, H., Isobe, T., Uchiyama, S., Miyake, K., et al. (2016). Structural Analysis Reveals that Toll-like Receptor 7 Is a Dual Receptor for Guanosine and Single-Stranded RNA. Immunity *45*, 737–748. https://doi.org/10.1016/j.immuni.2016.09.011.

48. Tangye, S.G., Al-Herz, W., Bousfiha, A., Cunningham-Rundles, C., Franco, J.L., Holland, S.M., Klein, C., Morio, T., Oksenhendler, E., Picard, C., et al. (2022). Human Inborn Errors of Immunity: 2022 Update on the Classification from the International Union of Immunological Societies Expert Committee. J. Clin. Immunol., 1–35. https://doi.org/10.1007/s10875-022-01289-3.

49. Cruz, R., Diz-de Almeida, S., López de Heredia, M., Quintela, I., Ceballos, F.C., Pita, G., Lorenzo-Salazar, J.M., González-Montelongo, R., Gago-Domínguez, M., Sevilla Porras, M., et al. (2022). Novel genes and sex differences in COVID-19 severity. Hum. Mol. Genet. *31*, 3789–3806. https://doi.org/10.1093/hmg/ddac132.

50. Nakanishi, T., Pigazzini, S., Degenhardt, F., Cordioli, M., Butler-Laporte, G., Maya-Miles, D., Bujanda, L., Bouysran, Y., Niemi, M.E.K., Palom, A., et al. Age-dependent impact of the major common genetic risk factor for COVID-19 on severity and mortality. J. Clin. Invest. *131*, e152386. https://doi.org/10.1172/JCI152386.

51. Brown, G.J., Cañete, P.F., Wang, H., Medhavy, A., Bones, J., Roco, J.A., He, Y., Qin, Y., Cappello, J., Ellyard, J.I., et al. (2022). TLR7 gain-of-function genetic variation causes human lupus. Nature *605*, 349–356. https://doi.org/10.1038/s41586-022-04642-z.

52. Fricke-Galindo, I., Martínez-Morales, A., Chávez-Galán, L., Ocaña-Guzmán, R., Buendía-Roldán, I., Pérez-Rubio, G., Hernández-Zenteno, R. de J., Verónica-Aguilar, A., Alarcón-Dionet, A., Aguilar-Duran, H., et al. (2022). IFNAR2 relevance in the clinical outcome of individuals with severe COVID-19. Front. Immunol. *13*, 949413. https://doi.org/10.3389/fimmu.2022.949413.

53. Muñiz-Banciella, M.G., Albaiceta, G.M., Amado-Rodríguez, L., Del Riego, E.S., Alonso, I.L., López-Martínez, C., Martín-Vicente, P., García-Clemente, M., Hermida-Valverde, T., Enríquez-Rodriguez, A.I., et al. (2023). Age-dependent effect of the IFIH1/MDA5 gene variants on the risk of critical COVID-19. Immunogenetics *75*, 91–98. https://doi.org/10.1007/s00251-022-01281-6.

54. O'Connor, T.D., Kiezun, A., Bamshad, M., Rich, S.S., Smith, J.D., Turner, E., NHLBIGO Exome Sequencing Project, ESP Population Genetics, Statistical Analysis Working Group, Leal, S.M., and Akey, J.M. (2013). Fine-scale patterns of population stratification confound rare variant association tests. PloS One *8*, e65834. https://doi.org/10.1371/journal.pone.0065834.

55. Gallo, G., Calvez, V., and Savoia, C. (2022). Hypertension and COVID-19: Current Evidence and Perspectives. High Blood Press. Cardiovasc. Prev. Off. J. Ital. Soc. Hypertens. *29*, 115–123. https://doi.org/10.1007/s40292-022-00506-9.

56. Schmidt, A., Röner, S., Mai, K., Klinkhammer, H., Kircher, M., and Ludwig, K.U. (2023). Predicting the pathogenicity of missense variants using features derived from AlphaFold2. Bioinforma. Oxf. Engl. *39*, btad280. https://doi.org/10.1093/bioinformatics/btad280.

57. Gao, H., Hamp, T., Ede, J., Schraiber, J.G., McRae, J., Singer-Berk, M., Yang, Y., Dietrich, A.S.D., Fiziev, P.P., Kuderna, L.F.K., et al. (2023). The landscape of tolerated genetic variation in humans and primates. Science *380*, eabn8153. https://doi.org/10.1126/science.abn8197.

58. Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., Pritzel, A., Wong, L.H., Zielinski, M., Sargeant, T., et al. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science *381*, eadg7492. https://doi.org/10.1126/science.adg7492.

Journal Pre-proof

Journal Pre-proof

Figure Titles and Legends

Figure 1: Study design and cohort characteristics.

A Candidate genes included in targeted sequencing, grouped according to source of evidence (details in Table S1). Genes known to cause Human Inborn Errors of Immunity[48] are highlighted in bold, and genes excluded during quality control due to low sequencing coverage are crossed out. **B** Workflow describing the main steps of sample preparation, genotyping, sequencing, and computational processing. Boxes colored in gray indicate steps that were performed in previous studies[25,26], MIP: Molecular Inversion Probe, QC: Quality control, PCA: Principal component analysis. **C** Number of individuals in the Italian (left) and Spanish (right) subcohorts. The number of COV_{hosp} individuals with no reported risk factors (as described in Table 1) is highlighted in red. The proportion of females is shown in parentheses. **D** Number of variants observed in the cohort in relation to their minor allele frequency (MAF). In the present study, variants with MAF<1% were denoted as rare variants, while all others were considered common. Intensity of color shading indicates whether (dark) or not (light) variants have been reported with allele frequency in gnomAD r2.1 exomes.

Journal Pre-proof

Figure 2: Association analysis for individual variants.

P-values (y-axis, negative log10) obtained in the association analysis of 1,211 non-singleton variants from the POP_{all} analysis. Variants are grouped according to the genes (x-axis, sorted alphabetically) in which they are located. Results for case-control definitions other than POP_{all} are provided in Figure S3. Dotted line: Lenient significance threshold, correcting for the number of variants tested ($\alpha=4.1x10^{-5}$). Dashed line: Strict significance threshold, also taking into account multiple testing due to additional case-control definitions ($\alpha=6.7x10^{-6}$). Variants with p-values below the lenient significance threshold are marked in green and were only found in genes selected based on prior GWAS evidence, i.e., *FYCO1* and *XCR1* at 3p21.31, *IFNAR2* at 21q22.11.

Journal Pre-proof

Figure 3: Results of the gene-based collapsing analysis for rare variants with MAF <0.1%.

P-values (y-axis, negative log10) are plotted for 52 genes (x-axis, sorted alphabetically). The various case-control definitions (see Table 1; excluding sex-stratified analyses) are depicted as symbols, while variant deleteriousness classes are coded according to color (M1: pLoF, M3 & M4: pLoF and moderate effect variants including missense in two graduations, C10+M1: CADD>10 or pLoF, SYN: synonymous, see Methods). Dashed line: Strict significance threshold, correcting for all tests conducted: (α=8.7x10^{-6}). Dotted line: More lenient significance threshold, correcting for the number of genes tested (α=9.6x10^{-4}). Results for sex-stratified analyses and variants with MAF<1% are provided in Figure S4.

Journal Pre-proof

Figure 4: Forest plot for *TLR7* rare variant gene burden according to variant classification.

A Odds ratios (ORs) of collapsed variants in *TLR7* are shown for $POP_{lowrisk}$ at different minor allele frequency groups (MAF) and deleteriousness predictions (class). Within each group, results are presented for all individuals and for sex-stratified analyses. Error bars indicate 95% confidence intervals. SYN: synonymous, C10+M1: CADD>10 or pLoF, bLoF: biochemical evidence for a loss-of-function effect, 3D-P: variant class based on 3D protein structure, see Methods. SYN variants with MAF<0.001 were only present in controls (OR=0.0, no confidence interval calculable). **B** Presence of 3D-P *TLR7* (MAF<0.1%) variant carriers (black dots) in all COV_{hosp} individuals (gray blue), COV_{hosp} with age<60y (light blue) and COV_{hosp} with no reported risk factors ("no RF", dark blue). The number of individuals within each set is indicated by area, and is specified in the outer legend. Percentages in brackets represent carrier ratios.

Journal Pre-proof

Figure 5: Location of rare *TLR7* variants within TLR7 protein domains.

A Rare, deleterious *TLR7* variants ($POP_{lowrisk}$, C10+M1, MAF<0.1%) are mapped on the protein domains of TLR7 (x-axis: amino acid position). Phenotype, according to the $POP_{lowrisk}$ case-control definition, and the sex of variant carriers is indicated by color or caption. Variants of carriers previously reported in Asano et al.[19] (see Methods and Table S2) are indicated by asterisks (*). TLR7 domains: LRR-NT (leucine-rich repeat, N-terminal, aa. 27-65) orange; LRR regions 1-26 (aa. 66-786) blue; LRR-CT (leucine-rich repeat, C-terminal, aa. 787-839), violet; TIR (Toll/interleukin-1 receptor) domain (aa. 889-1033), green. **B** TLR7 dimer overview, interface highlighted as gray surface (also in panel C). **C** Non-synonymous variants from panel A are highlighted in the 3D conformation of one TLR7 subunit (PDB ID: 5GMH) and are presented from two angles. Phenotype ($POP_{lowrisk}$, see A) and sex of the variant carriers are indicated by color coding. Variants within 5 Ångström of the subunit interface are highlighted by a hash (#, also in panel A). Variants located downstream of position T858 could not be plotted due to absence of the respective residues from the structure. Visualized using PyMOL Molecular Graphics System (Version 2.5.5 Schrödinger, LLC).

Journal Pre-proof

Tables

Table 1: Case-control definitions used in the present study

Analysis	cases	N cases (females/ males)	controls	N controls (females/ males)
Case-control definitions for analyses involving population-based controls (POP)				
1) POP_{all}	Individuals hospitalized for COVID-19 who required respiratory support (COV_{hosp})	1,772 (605/1,167)	Individuals from the general population with unknown SARS-CoV-2/COVID-19 status (Population controls)	5,347 (2,102/ 3,245)
2) $POP_{lowrisk}$	COV_{hosp} with no reported risk factors*	378 (126/252)	same as above	5,347 (2,102/ 3,245)
Case-control definitions for analyses involving COVID-19 hospitalized individuals (COV_{hosp}) only				
3) COV_{hosp} by risk factors	COV_{hosp} with no reported risk factors*	378 (126/252)	COV_{hosp} with two or more of the reported risk factors*	726 (244/482)
4) COV_{hosp} by respiratory support	COV_{hosp} requiring respiratory support level 3 (intubation) or 4 (ECMO, highest level)	478 (115/363)	COV_{hosp} requiring respiratory support level 1 (oxygen mask only, lowest level)	661 (284/377)

*Risk factors for which phenotype data were broadly available: age ≥ 60 years, diabetes, hypertension, coronary artery disease (CAD). Notably, 66 of 1,772 COV_{hosp} individuals did not have sufficient information on comorbidities and were therefore excluded from the risk

Journal Pre-proof

factor based stratification ($POP_{lowrisk}$, COV_{hosp} by risk factors). ECMO=extracorporeal membrane oxygenation.

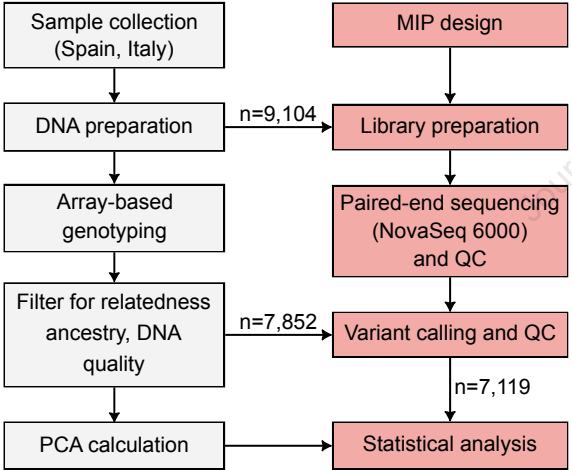
Journal Pre-proof

Journal Pre-proof

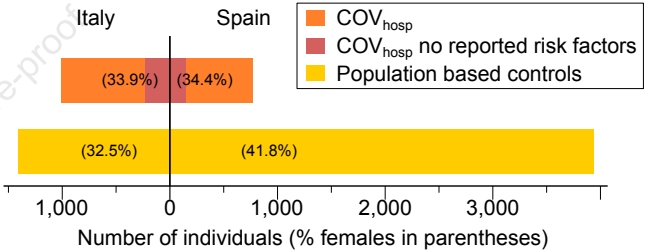
A

GWAS loci	Functional evidence			Diagnostics
ABO, CCR9, CDH18, CXCR6, FYCO1, IL10RB, LZTFL1, SLC6A20, XCR1	*ACE2, CCR3,* ***IFNAR1****,* ***IFNAR2****, TMPRSS2*	*ADAM10, ADAM17, CASP1,* ***CASP8****, CCL2, CCL20, ~~CCL3~~, CCR1, CCR2,* ***CD46****,* ***CFB****, ~~CFD~~,* ***CFH****,* ***CFI****, CTSB, CTSL, ~~CXCL1~~, CXCL10, CXCL3, CXCL8, DDX1, DDX58, FURIN,* ***IFIH1****, IFITM3,* ***IFNGR1****, IL1B, LY75, MBL2,* ***MYD88****,* ***NLRP3****,* ***RIPK1****, RIPK3,* ***TLR3****, TLR4, TNF*	***TLR7***	*AP3B1, SIGLEC1,* ***TBK1****, TNFRSF13B*

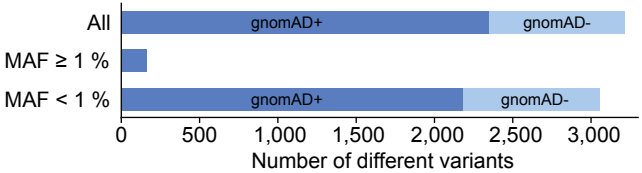
B

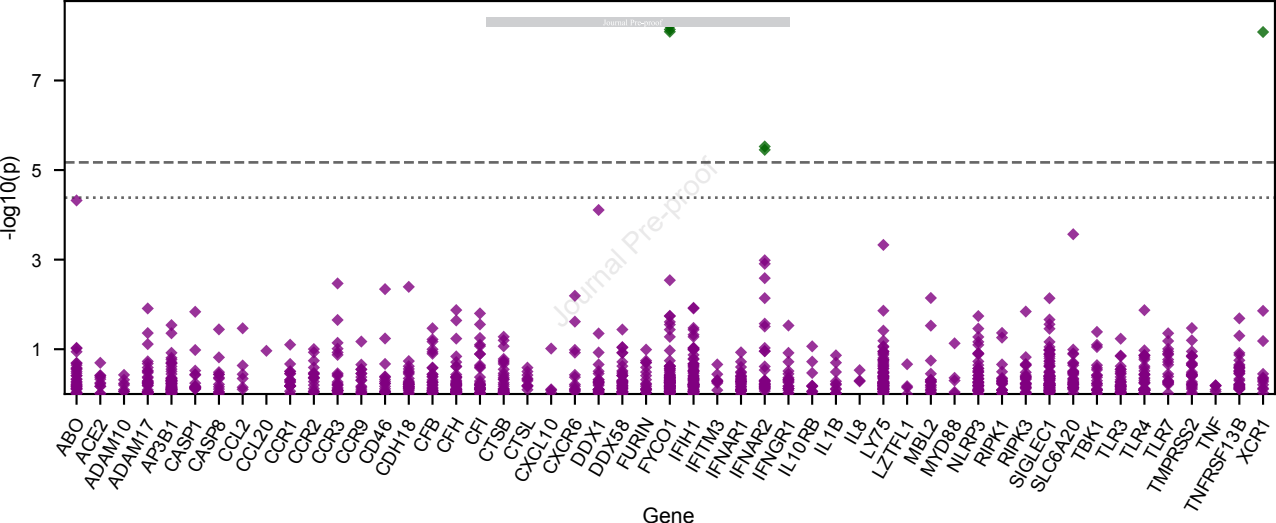


C

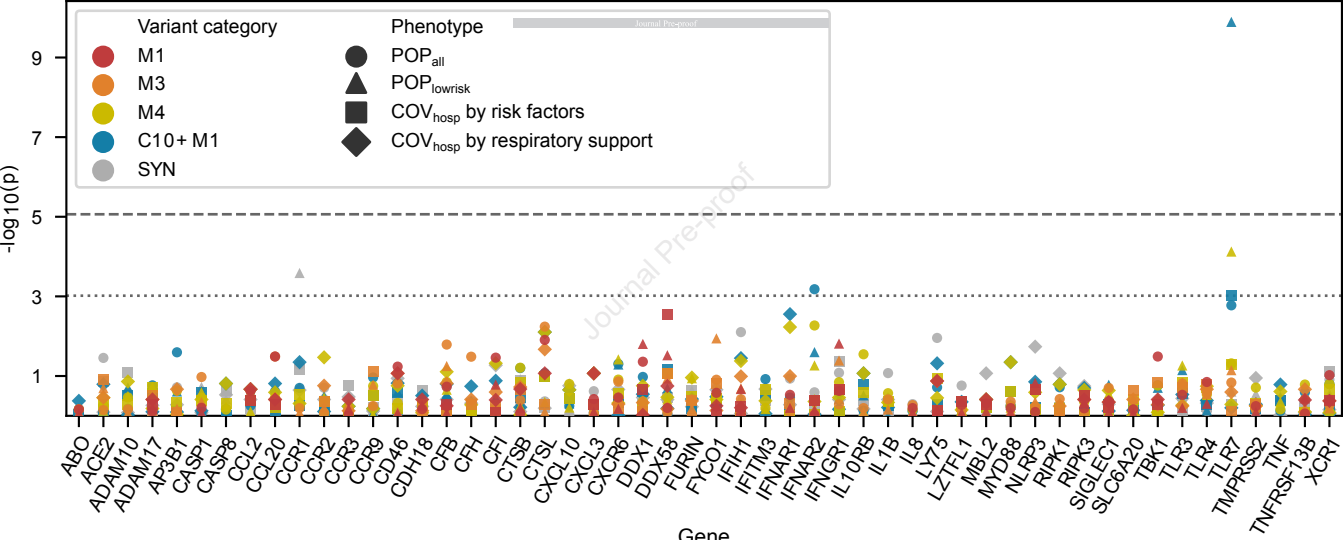


D



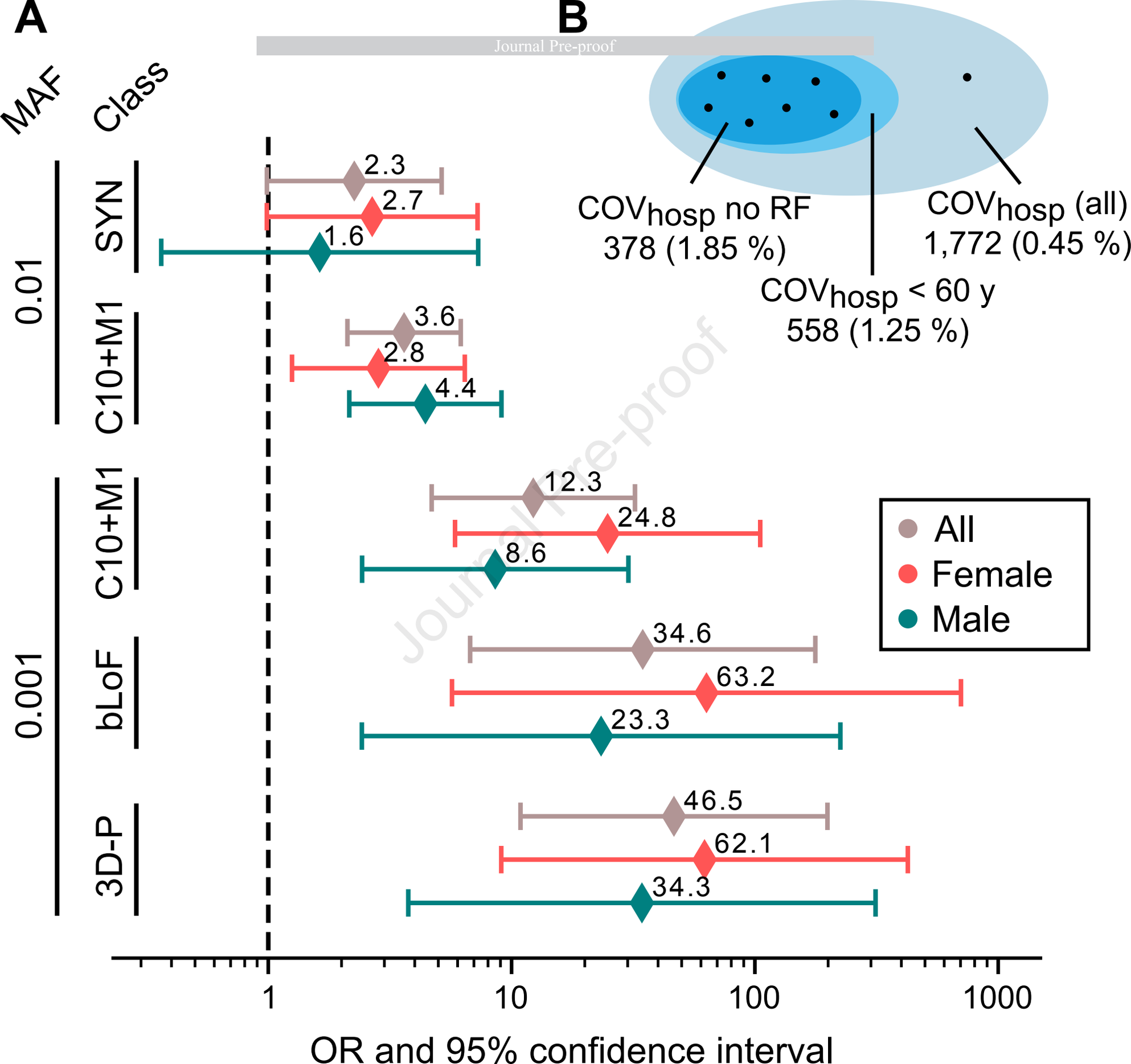


Journal Pre-proof

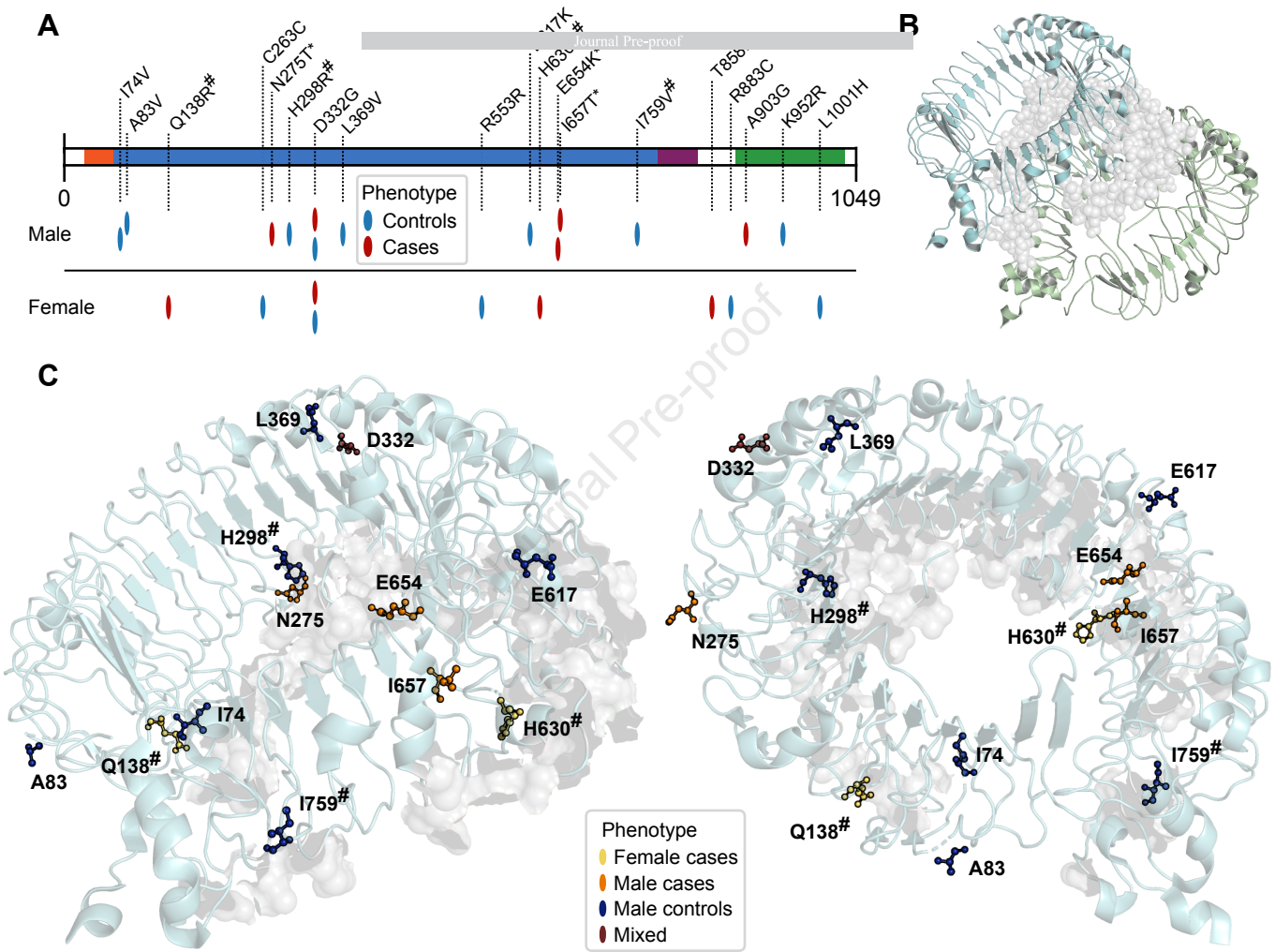


Journal Pre-proof

Journal Pre-proof



Journal Pre-proof



Journal Pre-proof

This study explores genetic factors in severe COVID-19 in a Spanish/Italian cohort. Rare deleterious *TLR7* variants were found in 2.4% of young cases. Incorporation of functional assay or protein modeling data yielded increased effect sizes. Further, *TLR7* signals in females suggest additional mechanisms beyond X-chromosomal recessive inheritance.

Journal Pre-proof