# Genetic determinants of respiratory diseases and their clinical implications

**Tomoko Nakanishi**

Kyoto-McGill International Collaborative Program in Genomic Medicine

Department of Human Genetics, Faculty of Medicine,

McGill University, Montreal

Department of Respiratory Medicine, Graduate School of Medicine,

Kyoto University, Kyoto

February 2022

A thesis submitted to McGill University and Kyoto University in partial fulfillment of the requirements of the degree of Doctor of Philosophy



McGill


京都大学
KYOTO UNIVERSITY

## Abstract

Despite the successful prosecution of large-scale genetic association studies to identify thousands of genetic determinants for respiratory diseases, the way to translate these insights into clinical fields has remained unrefined. This doctoral thesis presented potential avenues to translate such genetic findings into the clinical management of respiratory diseases, encompassing alpha-1 antitrypsin deficiency (AATD), idiopathic pulmonary fibrosis (IPF) and coronavirus disease 2019 (COVID-19).

First, AATD is a rare monogenic disorder caused by mutations in the *SERPINA1* gene. We demonstrated that in the UK Biobank, among 140 European-ancestry participants with the PI*ZZ genotype of *SERPINA1*, the most common AATD-associated genotype, only nine were diagnosed with AATD. Nonetheless, those with PI*ZZ had a substantially increased burden of multiple symptoms and diseases, including COPD and cirrhosis. It indicates that genetic testing would help identify those at risk and contribute to early intervention, such as smoking cessation counselling.

Second, IPF is a progressive, fatal fibrotic form of interstitial lung disease leading to decreased lung compliance and resulting in respiratory failure. We used a Mendelian Randomization (MR) approach, a causal inference technique, to efficiently scan hundreds of plasma proteins to identify proteins which may play causal roles in IPF susceptibility. We identified that circulating FUT3 was associated with a reduced risk of IPF (odds ratio [OR]: 0.81 per 1 SD increase in FUT3). FUT3 could be further investigated as drug targets for treatment, as well as noninvasive biomarkers of disease risk.

Lastly, we applied the same approach to study the novel COVID-19 pandemic. We evaluated the major common genetic risk for severe COVID-19 on chromosome 3, which was tagged by the rs10490770 C allele. Risk allele carriers age ≤60 years had higher odds of death or severe respiratory failure (OR: 2.7) compared with those >60 years (OR: 1.5). This risk variant improved the prediction of severe disease similarly to most clinical risk factors. Thus, it implicates the use of this genetic risk to realize genetics-guided clinical management. Similarly, we also used MR to identify proteins which could influence COVID-19 severity and susceptibility. It identified that an SD increase in OAS1 levels was associated with reduced COVID-19 death or ventilation needs (OR: 0.54), hospitalization (OR: 0.61), and susceptibility (OR: 0.78). Known pharmacological agents that increase OAS1 levels could be explored for their effect on COVID-19 outcomes.

In summary, this doctoral thesis provided a novel contribution to the field of genetics in respiratory medicine, by demonstrating potential opportunity to realize clinical benefits of emerging worldwide genomic efforts and by identifying potentially druggable disease-influencing plasma proteins.

**Résumé**

Les études d'association génétique à grande échelle ont permis d'identifier avec succès des milliers de déterminants génétiques des maladies respiratoires, cependant la manière de traduire ces connaissances dans les domaines cliniques n'a pas encore été clairement définie. Cette thèse de doctorat présente des pistes potentielles pour traduire ces découvertes génétiques dans la gestion clinique des maladies respiratoires, englobant le déficit en alpha-1-antitrypsine (AATD), la fibrose pulmonaire idiopathique (FPI) et la maladie à coronavirus 2019 (COVID-19).

Premièrement, l'AATD est un trouble monogénique rare causé par des mutations du gène *SERPINA1*. Nous avons démontré que parmi 140 participants d'ascendance européenne de la « UK Biobank » (UKB) présentant le génotype PI*ZZ de *SERPINA1*, le génotype le plus communément associé à l'AATD, seuls neuf avaient reçu un diagnostic d'AATD. Cependant, ceux ayant ce génotype présentaient de nombreux symptômes et maladies, dont la BPCO et la cirrhose. Ces résultats demontrent que des tests génétiques permettraient d'identifier les personnes à risque et de contribuer à une intervention précoce, telle que des conseils pour l'arrêt du tabac.

Deuxièmement, la FPI est une forme fibrotique progressive et mortelle de maladie pulmonaire interstitielle qui entraîne une diminution de la compliance pulmonaire et une insuffisance respiratoire. Nous avons utilisé une approche de randomisation mendélienne (RM), une technique d'inférence causale, pour analyser efficacement des centaines de protéines plasmatiques afin d'identifier les protéines susceptibles d'influencer cette pathologie. Nous avons identifié que la protéine FUT3 en circulation était associée à un risque réduit de FPI (odds ratio [OR] : 0,81 pour une augmentation de 1 ecart-type de FUT3). Ainsi, la protéine

FUT3 pourrait faire l'objet de recherches plus approfondies en tant que cible thérapeutique et en tant que biomarqueur non invasif du risque de FPI.

Enfin, nous avons appliqué la même approche pour étudier la nouvelle pandémie de la COVID-19. Nous avons évalué le principal risque génétique commun sur le chromosome 3 de la COVID-19 sévère, marqué par l'allèle C rs10490770. Les porteurs de cet allèle âgés de ≤60 ans avaient un risque plus élevé de décès ou d'insuffisance respiratoire sévère (odds ratio [OR]: 2,7) par rapport à ceux âgés de >60 ans (OR: 1,5). Ce facteur de risque a amélioré la prédiction de la forme sévère de la maladie de la même manière que la plupart des facteurs de risque cliniques, démontrant l'utilité de réaliser une gestion clinique guidée par la génétique. Nous avons également utilisé la RM pour identifier les potentielles protéines influençant la sévérité et la susceptibilité à la COVID-19. Nous avons mis en évidence qu'une augmentation de l'ecart-type des taux d'OAS1 était associée à une réduction de la mortalité ou des besoins de ventilation dus à la COVID-19 (OR : 0,54), de l'hospitalisation (OR : 0,61) et de la susceptibilité (OR : 0,78). Les agents pharmacologiques connus qui augmentent les taux d'OAS1 pourraient ainsi être explorés pour leurs effets sur la COVID-19.

En résumé, cette thèse de doctorat a apporté une nouvelle contribution au domaine de la génétique en médecine respiratoire, en démontrant l'opportunité potentielle de tirer des bénéfices cliniques des efforts émergents en génomique à l'échelle mondiale et en identifiant de potentielles protéines plasmatiques, influençant les maladies, comme cibles thérapeutiques.

**Table of Contents**

## List of Abbreviations

**1000G:** 1000 Genomes Project

**AATD:** Alpha-1 antitrypsin deficiency

**AIC**: Akaike information criterion

**AKI**: Acute kidney injury

**AMI**: Acute myocardial infarction

**ARDS**: Acute respiratory distress syndrome

**AUC**: Area under the curve

**BBJ**: Biobank Japan

**BF:** Bayes factor

**BH:** Benjamini-Hochberg

**BIC**: Bayesian information criterion

**BMI**: Body Mass Index

**Bp:** Base pair

**BQC19**: Banque québécoise de la COVID-19

**CA19-9**: Carbohydrate antigen 19-9

**CDC**: Centers for Disease Control and Prevention

**CEA**: Carcinoembryonic antigen

**CI**: Confidence Interval

**CIHR**: Canadian Institutes of Health Research

**CLPP**: colocalisation joint posterior probability that the variants are causal for two traits calculated by eCAVIAR

**COPD**: Chronic Obstructive Pulmonary Diseases

**COVID-19**: Coronavirus Disease 2019

**CHUM**: the Centre Hospitalier de l'Université de Montréal

**DM**: Diabetes mellitus

**DNA**: Deoxyribonucleic Acid

**DVT**: Deep venous thrombosis

**gnomAD**: Genome Aggregation Database

**eGFR**: Estimated glomerular filtration rate

**eQTL**: Expression quantitative trait locus

**$FEV_1$**: Forced Expiratory Volume in 1 second

**FIMM**: Institute for Molecular Medicine Finland

**FRQS**: Fonds de recherche du Québec – Santé

**FVC**: Forced Vital Capacity

**GWAS**: Genome-Wide Association Study

**HR**: hazard ratio

**HWE**: Hardy-Weinberg Equilibrium

**ICD10**: International Classification of Diseases, 10th Revision

**ICU**: intensive care unit

**ILD**: Interstitial lung disease

**IPF**: Idiopathic pulmonary fibrosis

**IVW**: Inverse variance weighted

**JGH**: the Jewish General Hospital

**LD**: Linkage Disequilibrium

**MAF**: Minor allele frequency

**Mbp**: Megabase pairs

**MR**: Mendelian randomization

**NRI**: net reclassification improvement

**OPCS**: Office of Population Censuses and Surveys

**OMIM**: Online Mendelian Inheritance in Man

**OR**: odds ratio

**PC**: principal component

**PE**: pulmonary embolism

**PFT**: pulmonary function testing

**PheWAS**: phenome-wide association study

**PP4**: posterior probability that the two traits share causal variants calculated by the coloc R package

**pQTL**: protein quantitative trait locus

**PRS** : polygenic risk score

**QQ-plot**: Quantile-quantile plot

**RFU**: relative fluorescent unit

**SARS-CoV2**: severe acute respiratory syndrome coronavirus 2

**SD**: standard deviation

**SE**: standard error

**SOMAmer**: Slow Off-Rate Modified Aptamers

**SNP**: single nucleotide polymorphism

**sQTL**: Splicing quantitative trait locus

**UIP**: Usual interstitial pneumonia

**UKB**: UK Biobank

**ULN**: upper limit of normal

**UMAP**: uniform manifold approximation and projection.

**URL**: Universal Resource Locator

**VEP**: Variant Effect Predictor

**VTE**: Venous Thromboembolism

**WGS**: Whole Genome Sequencing

**WHO**: World Health Organization

**WT**: Wild-type

## List of Figures

**Chapter 2**



**Chapter 3**



**Chapter 4**



**Chapter 5**

## List of Tables

**Chapter 2**



**Chapter 3**



**Chapter 4**



**Chapter 5**

**Acknowledgments**

This doctoral thesis would not have been possible without the amazing support I received from some exceptional individuals and institutions.

First, I would like to begin by thanking my supervisor, Dr. Brent Richards, for providing me with a opportunity to grow into the researcher I am today. I came to Montreal in 2019 as an international PhD student of McGill-Kyoto International Collaborative Program in Genomic Medicine. Thank you for everything what you have done for me, Brent. Although I had very little exposure to research and very few publications at the beginning, I remember, when we first met in your office, you said « I will guide you to become a researcher who deserves a principal investigator (PI) anywhere in the world. ». I thought you were joking, but you were very serious and you convinced me. Your words indeed kept me motivated during the past three years of my PhD. I completed my graduate courses successfully, published over a dozen manuscripts, won more awards than I ever thought before, and most importantly, started developing skills to become an independent PI. Your supervision has thus positioned me for future success in academia. All of my accomplishment wouldn't be possible without you, so thank you.

Next, I would like to thank my co-supervisor, Dr. Toyohiro Hirai. Your supervision, as a clinician-scientist in respirology, largely supported my PhD research. At the beginning when I decided to join the Department of Respiratory Medicine at Kyoto University, I never thought of joining this joint-PhD program. I understand it was not an easy decision to allow me to study in Canada during my PhD. You kindly respect my motivation to focus on research and to study abroad, which importantly supported my success in my PhD.

I would like to thank two Directors of the McGill-Kyoto joint-PhD program; Dr. Fumihiko Matsuda and Dr. G. Mark Lathrop. This joint-PhD program was initiated and has been kept maintained owing to your tremendous effort. Both of you also supported my carrier personally. All of my PhD works were impossible without your support, so thank you.

I would like to thank my supervisory committee members: Drs. Celia Greenwood, Deborah Assayag, and Simon Gravel, whose doors were always open for discussion. They all have been important additional sources of knowledge for my research. I particularly appreciated that at the last committee meeting, my committee members spent additional 30 mins to discuss my future opportunity and my carrier building. Celia, I learned a lot from your course works, and from the discussion with you, especially in the long-COVID project. Deborah, I appreciate your scientific inputs as an ILD specialist, which helped me a lot in the IPF proteome MR project. Simon, I like your passion and clear motivation to understand population genetics through mathematics. One of the largest regrets from my PhD is that I wanted to learn from you and work with you more than I did. Thank you all, for your supervision from the different expertise.

To my collaborators in the COVID-19 Host Genetics Initiative (COVID-19 HGI), I appreciate all of your team effort to tackle the worldwide crisis. To Dr. Andrea Ganna, who led the COVID-19 HGI and who is the senior author of one of my manuscripts, thank you. Owing to your mentorship, I could remotely be a visiting researcher at your lab, Institute for Molecular Medicine Finland, and remotely worked with many of your trainees. Andrea, thank you for your guidance for all of my work in COVID-19 HGI and thank you for allowing me to lead the large collaborative study that became Chapter 3 of my thesis. Despite I was physically in Japan and Canada, our communication was as smooth as I felt I were in Finland while I worked with you.

To my collaborators in the field of respirology in the UK and the US, thank you for sharing your brilliance and for collaborating on Chapters 2 and 4. Dr. William Cookson, Dr. David Schwartz, Dr. Paul Wolters, Dr. Louise Wain, Dr. Gisli Jenkins, and Dr. Richard Allen, your expertise in respirology and genetics is what made a lot of our work possible.

During my three years at McGill, I have come to know several members of our lab. Vince, you have been always my big helper whenever I was in an impasse in my analyses. Without your help and support, I wouldn't be sitting here writing these acknowledgments. You have been an invaluable mentor during my graduate studies, so thank you. Guillaume, we worked very closely in COVID-19 research. I was impressed by your capability of writing in a fast and clear manner. I learned a lot from you how to communicate with collaborators respectfully but with a sense of humor. You were one of the best academic partners for me toward the same COVID-19 project. Thank you. Sirui, we spent countless hours discussing research and a lot that had nothing to do with research. Your support was tremendously appreciated in many aspects of my stay in Canada. Laetitia, thank you for translating my English into French for many occasions. I know it was cumbersome to you, but you always kindly helped me. There have been many others that I would like to thank: Despoina, Agustin, John, Julyan, Haoyu, Dave, Yossi, Yiheng, and Kevin. Thank you.

During my four years at Kyoto, I also have come to know several members of the ILD group. Dr. Handa, you have been always generous and respectful with my motivation to focus on research and to study abroad, which importantly supported my success in my PhD. There have been many others that I would like to thank: Dr. Tanizawa, Dr. Nakatsuka, Dr. Ikezoe, Dr. Ikegami, Dr. Niwamoto, Dr. Mori, and Dr. Yamada. Thank you.

To my parents, both of you have influenced and encouraged me in incredible ways. Both of you, as role models of an independent researcher in molecular biology and pediatric neurology, respectively, were important sources of knowledge for my research. You also supported me as a family member remotely during the lockdown in Montréal and closely during my stay in Japan. Without your support, I do not think I could finish my tough PhD journey, so thank you.

Finally, I would like to thank my sources of funding, without which none of this work would have been possible. The Japan Society for the Promotion of Science, McGill University, Kyoto University, and the Lady Davis Institute have all provided me with stipend funding. The Japan Society for the Promotion of Science, various faculties, divisions, or departments of McGill and Kyoto have all provided me with awards. Each award was accepted with humility and gratitude, and I thank them.

## Contribution to original knowledge

This doctoral thesis sought clinical implications of genetic determinants of respiratory diseases by providing the potential avenue in genomics-guided clinical management (Chapters 2 and 3) and identifying potentially druggable plasma proteins which influence disease susceptibility and severity through Mendelian randomization (MR) (Chapters 4 and 5).

Chapter 2 is titled "The undiagnosed disease burden associated with alpha-1 antitrypsin deficiency genotypes". I found that in UK Biobank (UKB), among 140 European-ancestry participants with PI*ZZ genotype of *SERPINA1*; the most common AATD-associated genotype, only nine were diagnosed with AATD. Nonetheless, those with PI*ZZ had a substantially increased burden of multiple symptoms and diseases, including COPD and cirrhosis. Genetic testing could help to identify these at-risk individuals. Chapter 3 is titled "Age-dependent impact of the major common genetic risk factor for COVID-19 on severity and mortality". In this study, I combined and harmonized individual-level data from 13,888 COVID-19 patients from 17 cohorts in 9 countries to assess the association of the major common COVID-19 genetic risk factor (chromosome 3 locus tagged by rs10490770) with COVID-19 severity outcomes and COVID-19-related complications. Risk allele carriers had increased odds of several COVID-19 complications: not only severe respiratory failure (OR, 2.1; 95% CI, 1.6–2.6), but also venous thromboembolism (OR, 1.7; 95% CI, 1.2–2.4), and hepatic injury (OR, 1.5; 95% CI, 1.2–2.0). Risk allele carriers aged 60 years and younger had higher odds of death or severe respiratory failure (OR, 2.7; 95% CI, 1.8–3.9) compared with those more than 60 years old (OR, 1.5; 95% CI, 1.2–1.8; interaction, P = 0.038). This risk variant improved the prediction of death or severe respiratory failure similarly to, or better

than, most established clinical risk factors. In both manuscripts, we provided a potential avenue to use genetic testing to identify those at risk for early intervention.

In Chapters 4 and 5, we applied the Mendelian randomization (MR) approach to the recent genome-wide association studies (GWASs) to identify potential disease-influencing plasma proteins for respiratory diseases. Chapter 4 is entitled "Genetically increased circulating FUT3 level leads to reduced risk of Idiopathic Pulmonary Fibrosis: a Mendelian Randomisation Study". We identified that circulating FUT3 was associated with a reduced risk of idiopathic pulmonary fibrosis (IPF) (OR: 0.81 per 1 SD increase in FUT3). Chapter 5 is entitled "A Neanderthal OAS1 isoform protects individuals of European ancestry against COVID-19 susceptibility and severity". We identified that SD increase in OAS1 levels was associated with reduced COVID-19 death or ventilation (OR: 0.54), hospitalization (OR: 0.61), and susceptibility (OR: 0.78). Both plasma proteins could be further investigated as drug targets for treatment, as well as noninvasive biomarkers of disease risk.

In summary, this doctoral thesis provides novel clinical implications of genetic determinants of a variety of respiratory diseases, by showcasing the potential avenue for genomics-guided clinical management and by identifying potential disease-influencing plasma proteins that could be further investigated as therapeutic targets and biomarkers.

## Format of the Thesis

This is a manuscript-based thesis format as described in the Thesis Preparation Guidelines by the Department of Graduate and Postdoctoral Studies. This thesis contains seven chapters. Chapter 1 is an introduction to this thesis. Chapters 2 to 5 have been published in *the European Respiratory Journal*, *the Journal of Clinical Investigation*, *the European Respiratory Journal,* and *Nature Medicine*, respectively. Chapter 6 is a discussion of Chapters 2 to 5. Chapter 7 is a conclusion with future aims for Chapters 2 to 5. A summary of other publications can be found in the Appendix.

## Contribution of Authors

Chapter 2 is a manuscript authored by Tomoko Nakanishi, Vincenzo Forgetta, Tomohiro Handa, Toyohiro Hirai, Vincent Mooser, G. MarkLathrop, William O.C.M. Cookson and J. Brent Richards. It was published in *the European Respiratory Journal* on December 10th, 2020. Conception and design: T. Nakanishi and J.B. Richards. Data analyses: T. Nakanishi. Data acquisition: T. Nakanishi, V. Forgetta and J.B. Richards. Interpretation of data: T. Nakanishi, V. Forgetta, T. Handa, T. Hirai, V. Mooser, G.M. Lathrop, W.O.C.M. Cookson and J.B. Richards. Intellectual contribution to the manuscript: T. Nakanishi, V. Forgetta, T. Handa, T. Hirai, V. Mooser, G.M. Lathrop, W.O.C.M. Cookson and J.B. Richards. All authors were involved in preparation of the further draft of the manuscript and revising it critically for content. All authors gave final approval of the version to be published. T. Nakanishi and J.B. Richards are the guarantors. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. Specifically, as the first author of this paper, I curated data from UK Biobank and performed all the analyses. I wrote the abstract, introduction, methods, results and discussion, and all relevant tables and figures, both main and supplementary.

Chapter 3 is a manuscript authored by Tomoko Nakanishi, Sara Pigazzini, Frauke Degenhardt, Mattia Cordioli, Guillaume Butler-Laporte, Douglas Maya-Miles, Luis Bujanda, Youssef Bouysran, Mari E.K. Niemi, Adriana Palom, David Ellinghaus, Atlas Khan, Manuel Martínez-Bueno, Selina Rolker, Sara Amitrano, Luisa Roade Tato, Francesca Fava, FinnGen, The COVID-19 Host Genetics Initiative (HGI), Christoph D. Spinner, Daniele Prati, David Bernardo, Federico Garcia, Gilles Darcis, Israel Fernández-Cadenas, Jan Cato Holter, Jesus M. Banales, Robert Frithiof, Krzysztof Kiryluk, Stefano Duga, Rosanna Asselta, Alexandre C. Pereira, Manuel Romero-Gómez, Beatriz Nafría-Jiménez, Johannes R. Hov, Isabelle

Migeotte, Alessandra Renieri, Anna M. Planas, Kerstin U. Ludwig, Maria Buti, Souad Rahmouni, Marta E. Alarcón-Riquelme, Eva C. Schulte, Andre Franke, Tom H. Karlsen, Luca Valenti, Hugo Zeberg, J. Brent Richards, and Andrea Ganna. It was published in *the Journal of Clinical Investigation* on October 1, 2021. TN, GBL, BNJ, FG, RF, MRG, KUL, MB, S Rahmouni, MEAR, ECS, THK, LV, HZ, JBR, and AG conceived and designed the study. TN, SP, FD, MC, GBL, DMM, BNJ, YB, MEKN, DE, MMB, KUL, MEAR, LV, HZ, BR, and AG applied statistical, mathematical, computational, or other formal techniques to analyze or synthesize data. TN, FD, MC, GBL, DMM, BNJ, YB, MEKN, DE, MMB, S Rolker, SA, LRT, FF, CDS, FG, IFC, JCH, RF, RA, ACP, LB, JRH, IM, AR, KUL, MB, ECS, JBR, and AG curated data. TN, GBL, BNJ, S Rolker, RF, MRG, IM, KUL, MEAR, LV, HZ, BR, and AG interpreted data. DMM, SA, FF, CDS, DP, DB, FG, GD, JCH, RF, SD, MRG, JRH, IM, AR, KUL, MB, S Rahmouni, MEAR, ECS, THK, JBR, and AG acquired funding. TN, GBL, DMM, BNJ, YB, RF, IM, KUL, MEAR, JBR, and AG performed experiments. TN, GBL, MMB, MEAR, HZ, JBR, and AG developed or designed the methodology. TN, FD, DMM, S Rolker, CDS, DP, DB, FG, GD, JCH, JMB, JRH, IM, KUL, S Rahmouni, ECS, AF, THK, LV, JBR, and AG provided project administration. FG, GD, MRG, IM, S Rahmouni, MEAR, JBR, and AG provided resources. DMM, BNJ, FG, MRG, IM, KUL, S Rahmouni, MEAR, JBR, and AG supervised the experiments. TN, SP, FD, DE, AK, KK, and AG verified the overall replication/reproducibility of results as a separate activity. TN and AG prepared, created, and visualized the published work. TN, JBR, and AG wrote the original draft of the manuscript. TN, GBL, DMM, BNJ, AP, S Rolker, IFC, JCH, RF, KK, SD, RA, LB, JRH, IM, AR, AMP, KUL, MEAR, THK, LV, HZ, JBR, and AG reviewed and edited the manuscript. All authors were involved in further drafts of the manuscript and revised it critically for content. All authors gave final approval of the version to be published. The corresponding authors attest that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. Specifically, as the first

author of this paper, I curated and harmonized all individual-level data from all cohorts and performed all the analyses. I wrote the abstract, introduction, methods, results and discussion, and all relevant tables and figures, both main and supplementary.

Chapter 4 is a manuscript authored by Tomoko Nakanishi, Agustin Cerani, Vincenzo Forgetta, Sirui Zhou, Richard J. Allen, Olivia C. Leavy, Masaru Koido, Deborah Assayag, R. Gisli Jenkins, Louise V. Wain, Ivana V. Yang, G. Mark Lathrop, Paul J. Wolters, David A. Schwartz, J. Brent Richards. It was published online ahead of print in *the European Respiratory Journal* on Feburary, 2022. Conception and design: T. Nakanishi and J.B. Richards. Data analyses: T. Nakanishi and O.C. Leavy. Manuscript writing: T. Nakanishi and J.B. Richards. Data acquisition: R.J. Allen, R.G. Jenkins, L.V. Wain, P.J. Wolters and D.A. Schwartz. Interpretation of data: all authors. Intellectual contribution to the manuscript: all authors. All authors were involved in the preparation of the further draft of the manuscript and revising it critically for content. All authors gave final approval of the version to be published. T. Nakanishi and J.B. Richards are the guarantors. Specifically, as the first author of this paper, I performed all the analyses. I wrote the abstract, introduction, methods, results and discussion, and all relevant tables and figures, both main and supplementary.

Chapter 5 is a manuscript authored by Sirui Zhou, Guillaume Butler-Laporte , Tomoko Nakanishi, David R. Morrison, Jonathan Afilalo, Marc Afilalo, Laetitia Laurent, Maik Pietzner, Nicola Kerrison, Kaiqiong Zhao, Elsa Brunet-Ratnasingham, Danielle Henry, Nofar Kimchi, Zaman Afrasiabi, Nardin Rezk, Meriem Bouab, Louis Petitjean, Charlotte Guzman, Xiaoqing Xue, Chris Tselios, Branka Vulesevic, Olumide Adeleye, Tala Abdullah, Noor Almamlouk, Yiheng Chen, Michaël Chassé, Madeleine Durand, Clare Paterson, Johan Normark, Robert Frithiof, Miklós Lipcsey, Michael Hultström, Celia M. T. Greenwood, Hugo Zeberg, Claudia Langenberg, Elin Thysell , Michael Pollak, Vincent Mooser, Vincenzo

Forgetta, Daniel E. Kaufmann and J. Brent Richards. It was published in *Nature Medicine* on February 25, 2021. Conception and design: S.Z., G.B.L. and J.B.R. Data analyses: S.Z. and T.N. Data acquisition: T.N., G.B.L., D.M., D.E.K., J.A., M.A., L.L., E.B.R., D.H., N.K., Z.A., N.R., M.B., L.P., C.G., X.X., C.T., B.V., O.A., T.A., N.A., M.C., M.D., V.F., D.E.K. and J.B.R. Interpretation of data: S.Z., G.B.L., T.N., M.P., Y.C., D.E.K., V.F. and J.B.R. Funding acquisition: D.M., V.M., V.F. and J.B.R. Methodology: S.Z., K.Z., C.M.T.G. and J.B.R. Project administration: D.M., V.F. and J.B.R. Validation: S.Z., T.N., M.P., N.K., M.P., J.N., E.T., C.L., D.E.K. and J.B.R. Visualization: S.Z., T.N. and V.F. Writing—original draft: S.Z., G.B.L., T.N. and J.B.R. Writing—review and editing: S.Z., G.B.L., T.N., M.P., H.Z., V.M., M.P., R.F., M.L., M.H., C.P., D.E.K. and J.B.R. All authors were involved in further drafts of the manuscript and revised it critically for content. All authors gave final approval of the version to be published. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. Specifically, as the first author of this paper, I contributed to the data generation and curation of COVID-19 HGI (5.6.2) and BQC19 data (5.6.7), co-localization studies (5.4.2, 5.6.5), and sQTL and eQTL studies for OAS genes (5.4.4, 5.6.6). I wrote the above listed results (5.4.2, 5.4.4), the above listed methods (5.6.2, 5.6.5, 5.6.6, 5.6.7), discussion (5.5), and relevant tables (table 2) and supplementary figures.

# Chapter 1 : General introduction

## 1.1 Global impact of respiratory diseases

Humans cannot live without lungs, which are responsible for bringing oxygen into the body and helping get rid of waste gases with every exhale. The lung is one of the most important vital organs, which is vulnerable to infection and injury from the external environment because of its constant exposure to particles, chemicals, and infectious organisms in ambient air.

Respiratory diseases impose an immense worldwide health burden. Chronic obstructive pulmonary disease (COPD) is one of the most common respiratory diseases in the general population and can lead to decreased health status, exercise capacity, morbidity, and mortality. Approximately 10 percent of individuals aged 40 years or older have COPD, although the prevalence varies between countries and increases with age(1). COPD is defined by airflow limitation that is incompletely reversible compared to asthma(2), and is the third leading cause of death, contributing to 5% of global mortalities(3). WHO estimates 65 million people have moderate to severe COPD, out of more than 3 million people die of the disease(3). Idiopathic pulmonary fibrosis (IPF) is another progressive, fatal fibrotic interstitial lung disease (ILD) that typically affects adults aged more than 65, leading to decreased lung compliance, disrupted gas exchange, and the resultant respiratory failure(4). The incidence rate of IPF is lower compared to COPD, i.e. about 3-9 cases per 100000 people per year(5). Nevertheless, IPF has a striking impact on global health as the median survival time from the diagnosis is 3 years, which is equivalent or worse compared to several forms of cancers(6).

Emerged in 2019, the coronavirus disease of 2019 (COVID-19) pandemic caused by infection with SARS-CoV-2. COVID-19 has resulted in an enormous health and economic burden

worldwide. COVID-19 has rapidly become one of the most impactful respiratory diseases, as more than 20 million individuals have been infected as of end of January 2022(7). COVID-19 has also changed our personal lives, where new social attitudes, such as community mask-wearing, social distancing, enhanced personal hygiene and reduced travel, have been widely accepted. One of the most remarkable features of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection is the variation in consequences, which range from asymptomatic to life-threatening, viral pneumonia, and acute respiratory distress syndrome(8).

### 1.2 Genetic evidence for respiratory diseases

Like all other human disorders, respiratory diseases occur as a result of interactions between genetic and environmental factors. For example, smoking is the leading risk factors for many respiratory diseases, such as COPD and IPF. The chronologically taken environmental/occupational history may disclose other important risk factors for COPD, such as exposure to fumes or organic/inorganic dusts. These exposures help to explain the 20 percent of patients with COPD (defined by lung function alone), while the 20 percent of patients who die from COPD had never smoked(9). Genetics could partially explain the residual variablity of the disease onset.

Alpha-1 antitrypsin deficiency (AATD) is a monogenic cause of COPD by mutations in the *SERPINA1* gene inherited with incomplete penetrance(10) and has a prevalence of 1 case per 3000 to 5000 people(11). Alpha-1 antitrypsin (AAT) is an endogenous inhibitor of the proteolytic enzyme elastase and a severe deficiency of AAT enhances the burden of neutrophil elastase in the lungs, leading to emphysema(12). Intrahepatic accumulation of non-secreted AAT also predisposes to liver diseases including cirrhosis and hepatic

carcinoma(12). The most common disease-associated mutation is denoted PI*Z (p.Glu342Lys), and PI*ZZ homozygotes account for the most common severe phenotype of AATD(12). The compound heterozygous genotype PI*SZ, where PI*S is another missense mutation (p.Val264Glu), is associated with a more mildly increased risk of emphysema in smokers(13). PI*MM refers to homozygosity for wild-type alleles.

On the other hand, more complex respiratory diseases, such as IPF, have more polygenetic architecture. Owing to the prior family-based genetic studies for familial pulmonary fibrosis (FPF)(14–16) and large-scale genome-wide association studies (GWASs) for IPF(17, 18), both rare (with a minor allele frequency of less than 0.1%) variants (i.e. in telomere-related genes and surfactant-associated protein genes) and common variants (i.e. in *MUC5B*, *DSP*, and telomere-related genes) have been discovered to predispose to IPF.

### 1.3 Host genetics contributing to the COVID-19 outcomes.

COVID-19 is a respiratory illness appearently caused by the SARS-CoV-2 infection. Although established host factors, such as increasing age, male sex, and higher BMI(19), correlate with disease severity, these risk factors alone do not explain all of the variability in disease severity observed across individuals. Indeed, severe cases were observed among young individuals without apparent previous pre-existing risk factors, and sometimes they clustered in families(20), suggesting that human genetics play an role for disease risk. In the recent study(20), rapid clinical whole-exome sequencing of the COVID-19 patients demonstrated the segregation in available family members with loss-of-function variants of the X-chromosomal *TLR7*. Recent large-scale meta genome-wide association studies (GWAS) for COVID-19 outcomes, which is a study design to identify common genetic determinants fo complex diseases, identified several loci associated with COVID-19 severity

and susceptiility(21–23). These included members of the Toll-like receptor group such as *TLR3* and *TLR7*, type I interferon receptors (Interferon Alpha And Beta Receptor Subunit 2, *IFNAR2*), Janus kinase 1 (*JAK1*) and tyrosine kinase 2 (*TYK2*), interferon-stimulated genes such as oligoadenylate synthetase 1 (*OAS1*) (21–24). One of the major genetic risks for COVID-19 severity resides in chromosome 3, which confers about doubled-risk of ICU addmission(22). This locus on chromosome 3p21 includes the putative SARS-CoV-2 coreceptors *LZTFL1*. Using chromosome conformation capture and gene-expression analysis, a recent study identified the gain-of-function SNP for *LZTFL1*, rs17713054G>A, as a probable causative variant conferring increased risk of respiratory failure with COVID-19(25).

### 1.4 Limited evidence in translating the genetic findings into clinical management of respiratory diseases

The capacity to undertake genome-wide association studies (GWAS) has resulted in spectacular advances in the understanding of the genetic basis of common diseases(26). Although compelling signals have been found, often highlighting previously unsuspected biology, few studies have carefully investigated the way to translate these insights into clinical fields.

There are two principal routes through which such translation might be realized(26). The first translational route lies through using genetic profiling of individual patterns of disease predisposition to develop more personalized approaches to disease management. The major limitation for most complex traits is that the variants identified so far explain only a small proportion of individual variation in disease risk. Therefore, except for those with highly penetrant rare genetic variants, genetic profiling provides limited information on disease risk

beyond that available from conventional risk factors. Consequently, the practical example of genetic testing is limited in specific fields, such as prenatal screening(27), increased monitoring or prophylactic surgery for variants associated with hereditary cancer syndromes (i.e. *BRCA1/2*)(28), or additional testing or interventions for variants associated with hypertrophic cardiomyopathy(29).

The second translational route is identification of therapeutic targets within causal pathways, leading to novel therapeutic agents for treatment and/or prevention. Identification of causal pathways could bolster efforts to identify biomarkers, allowing improved disease prediction and monitoring of disease progression and treatment response, as the previous study estimated that selecting genetically supported targets could double the success rate in clinical development(30).

In the field of respiratory disease, the ultimate objectives of genetic studies to translate genetic findings into clinical practice remained far away from the accomplishment. There were very few occations where genetic testing is recommended, i.e. when people present with respiratory symptoms and are suspected of some diseases, such as cystic fibrosis, COPD, and tuberous sclerosis, genetic testing may be recommended. Although there were many canonical examples in other fields, such as statins (treatment of dyslipidemia) for *HMGCR*, toclizumab (treatment of rheumatoid arthritis) for *IL6R*(30, 31), there has been little evidence of genomics-guided drug development in respiratory diseases.

## 1.5 Objectives and hypothesis

The two major objectives of the presented doctoral thesis were to:

1. Seek the implications of genomics-guided clinical management (Chapters 2 and 3).

2. Identify novel potentially causal proteins to disease susceptibility and severity, which could serve as drug targets and potential biomarkers (Chapters 4 and 5).

Recent successes in the identification of genetic determinants of respiratory diseases(17, 23, 32) have increased confidence that this information can be translated into clinically beneficial improvements in management. To do so, in this doctoral thesis, I sought the two potential routes listed above, through which such translation might be effective.

In the first objective, I hypothesized that the understanding of individual patterns of disease predisposition through genetic profiling may aid in developing more personalized approaches to clinical management. To answer this question, in Chapter 2, I first examined the frequency of the PI*ZZ genotype in individuals with and without diagnosed AATD from UK Biobank and assessed the associations of the genotypes with clinical outcomes and mortality. In Chapter 3, I combined and harmonized individual-level data from 13,888 COVID-19 patients from 17 cohorts in 9 countries to assess the association of the major common COVID-19 genetic risk factor (chromosome 3 locus tagged by rs10490770) with COVID-19 severity outcomes and COVID-19-related complications. Through this investigation, I provided the potential route to improve the prediction of severe COVID-19 using genetic information.

In the second objective, I hypothesized that Mendelian randomization (MR) with circulating proteins may identify specific proteins that play causal roles in lung diseases, which could be attractive targets for drug discovery and biomarkers. This is because circulating proteins are easy to measure from blood, are more stable than mRNA, but are still able to target specific genes. In Chapters 4 and 5, we used Mendelian randomization analyses as a hypothesis-generating tool, which efficiently scanned hundreds of circulating proteins, to identify

potential causal proteins as possible therapeutic targets for IPF and COVID-19 severity and susceptibility, respectively.

In summary, the scientific hypothesis of this Thesis was twofold: First, through the largescale epidemiological study, we could obtain the understanding of individual patterns of disease predisposition through genetic profiling, which may aid in developing more personalized approaches to clinical management. Second, by using the MR approach, we could identify novel potentially disease-influencing proteins which could serve as drug targets and potential biomarkers.

**Connecting Text: Bridge Between Chapter 1 and Chapter 2**

Chapter 2 describes a project, which sought the potential implications of genomics-guided clinical management in a monogenic disorder, alpha-1 antitrypsin deficiency (AATD). AATD remains an underdiagnosed condition despite initiatives developed to increase awareness(33). The World Health Organization recommends testing all COPD patients, and the European Respiratory Society and American Thoracic Society guidelines recommend testing all symptomatic adults with persistent airway obstruction, individuals with unexplained liver disease, and adults with necrotizing panniculitis or multisystemic vasculitis(34). Despite the well-established causal evidence of the *SERPINA1* gene on AATD, current guidelines for α1-antitrypsin deficiency (AATD) state that adult population screening should only be done in high-risk populations(35). Underdiagnosis of AATD is problematic, especially for primary care physicians who attend to most COPD patients, and they are usually the first point of contact of patients with health care providers. Electrical medical records are increasingly used in clinical research to enhance the knowledge about the management and progression of the disease based on real-life data. Database studies may help to understand the real clinical practice and to design public health strategies to improve the quality of care. The objective of this study was to describe the patterns of diagnosis of AATD and to understand the potential implications of genetic screening for the disease in the population-based database; UK Biobank.

# Chapter 2: The undiagnosed disease burden associated with alpha-1 antitrypsin deficiency genotypes

## 2.1 Title page

**The undiagnosed disease burden associated with alpha-1 antitrypsin deficiency genotypes**

Tomoko Nakanishi[1,2,3,4,5], Vincenzo Forgetta[2], Tomohiro Handa[6], Toyohiro Hirai[4], Vincent Mooser[1,7], G. MarkLathrop[8], William O.C.M. Cookson[9,10] and J. Brent Richards[1,2,11]

**Affiliations:** [1]Dept of Human Genetics, McGill University, Montréal, QC, Canada. [2]Centre for Clinical Epidemiology, Dept of Medicine, Lady Davis Institute for Medical Research, Jewish General Hospital, McGill University, Montréal, QC, Canada. [3]Kyoto–McGill International Collaborative School in Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan. [4]Dept of Respiratory Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan. [5]Research Fellow of Japan Society for the Promotion of Science, Tokyo, Japan. [6]Dept of Advanced Medicine for Respiratory Failure, Graduate School of Medicine, Kyoto University, Kyoto, Japan. [7]Canada Excellence Research Chair in Genomic Medicine, McGill University, Montréal, QC, Canada. [8]McGill University and Genome Québec Innovation Centre, Montréal, QC, Canada. [9]National Heart and Lung Institute, Imperial College London, London, UK. [10]Royal Brompton and Harefield NHS Foundation Trust, London, UK. [11]Division of Endocrinology, Depts of Medicine, Human Genetics, Epidemiology and Biostatistics, Jewish General Hospital, McGill University, Montréal, QC, Canada.

**Published in:** Eur Respir J 2020; 56 : 2001441 doi: 10.1183/13993003.01441-2020

## 2.2 Abstract

Alpha-1 antitrypsin deficiency (AATD), mainly due to the PI*ZZ genotype in *SERPINA1*, is one of the most common inherited diseases. Since it is associated with a high disease burden and partially prevented by smoking cessation, identification of PI*ZZ individuals through genotyping could improve health outcomes. We examined the frequency of the PI*ZZ genotype in individuals with and without diagnosed AATD from UK Biobank, and assessed the associations of the genotypes with clinical outcomes and mortality. A phenome-wide association study (PheWAS) was conducted to reveal disease associations with genotypes. A polygenic risk score (PRS) for forced expiratory volume in 1 s ($FEV_1$)/forced vital capacity (FVC) ratio was used to evaluate variable penetrance of PI*ZZ. Among 458164 European-ancestry participants in UK Biobank, 140 had the PI*ZZ genotype and only nine (6.4%, 95% CI 3.4–11.7%) of them were diagnosed with AATD. Those with PI*ZZ had a substantially higher odds of COPD (OR 8.8, 95% CI 5.8–13.3), asthma (OR 2.0, 95% CI 1.4–3.0), bronchiectasis (OR 7.3, 95%CI 3.2–16.8), pneumonia (OR 2.7, 95% CI 1.5–4.9) and cirrhosis (OR 7.8, 95% CI 2.5–24.6) diagnoses and a higher hazard of mortality (2.4, 95% CI 1.2–4.6), compared to PI*MM (wildtype) (n=398 424). These associations were stronger among smokers. PheWAS demonstrated associations with increased odds of empyema, pneumothorax, cachexia, polycythaemia, aneurysm and pancreatitis. Polygenic risk score and PI*ZZ were independently associated with $FEV_1$/FVC <0.7 (OR 1.4 per 1-SD change, 95% CI 1.4–1.5 and OR 4.5, 95% CI 3.0–6.9, respectively). The important underdiagnosis of AATD, whose outcomes are partially preventable through smoking cession, could be improved through genotype-guided diagnosis.

## 2.3 Introduction

Alpha-1 antitrypsin deficiency (AATD) is one of the most common inherited respiratory diseases in people of European descent [1]. Alpha-1 antitrypsin (AAT) is an inhibitor of the proteolytic enzyme elastase and a severe deficiency of AAT enhances the burden of neutrophil elastase in the lungs, leading to emphysema [2]. In addition, intrahepatic accumulation of nonsecreted AAT predisposes to liver diseases [2].

AATD is caused by mutations in the *SERPINA1* gene that result in changes in the electrophoretic mobility of the protein predispose to AATD with incomplete penetrance [3, 4]. The most common disease-associated mutation is denoted PI*Z (p.Glu342Lys) and PI*ZZ homozygotes account for the most common phenotype of AATD [2]. The compound heterozygous genotype PI*SZ, where PI*S is another missense mutation (p.Val264Glu), is associated with a more mildly increased risk of emphysema in smokers [5]. PI*MM refers to homozygosity for wild-type alleles.

AATD is often clinically diagnosed after the identification of COPD or liver disease in individuals with a family history, and the average age at diagnosis is ~45 years [6]. A previous report, using estimates of allele frequencies from the literature [7], but without direct genotyping, estimated that only 1068 of expected 305009 PI*ZZ and PI*SZ individuals had been included in an international AATD registry [8].

Given the partial efficacy of AATD-specific therapies [9] and the availability of smoking cessation counselling, early diagnosis of AATD could promote earlier intervention with smoking cessation therapies and allow for the identification of family members at high risk. Given recent announcements of UK ambitions to sequence 5 million individuals [10], there

may exist an opportunity to identify individuals with high-risk genotypes and put in place appropriate diagnostic programmes to reduce the burden of this disease.

Here we sought to understand the prevalence of *SERPINA1* genotype status in UK Biobank and assess the diagnosis rate of AATD. We next explored the magnitude of association between *SERPINA1* genotypes and respiratory conditions, changes in spirometry results, other extrapulmonary conditions and all-cause mortality. Taking advantage of the large sample size of UK Biobank, we conducted a phenome-wide association study (PheWAS) to investigate potential associations of *SERPINA1* genotypes with other outcomes. Lastly, we calculated a polygenic risk score (PRS) for forced expiratory volume in 1 s ($FEV_1$)/ forced vital capacity (FVC) ratio to assess the interactions of *SERPINA1* genotypes and common variants affecting lung function.

## 2.4 Material and methods

### 2.4.1 UK Biobank study subjects

UK Biobank is a population-based cohort which recruited people aged 40–69 years from across the UK. We selected 458164 participants of European descent (defined in the supplementary material, figures S1 and S2, table S1) with nonmissing *SERPINA1* Z and S genotype information (rs28929474 and rs17580). Both of these variants were genotyped in UK Biobank, and therefore our study is not reliant upon imputation. The minor allele frequencies of rs28929474 and rs17580 were 0.020 and 0.048, respectively. The genotype definition and the quality control metrics of the genotypes are listed in supplementary table S2. The Z and S allele status of individuals of non-European descent is listed in supplementary table S3.

### 2.4.2 Ethical compliance

The UK Biobank was approved by the North West Multi-centre Research Ethics Committee and informed consent was obtained from all participants prior to participation.

### 2.4.3 Clinical data ascertainment

Prevalent disease was ascertained by self-reported physician-made diagnoses, self-reported recent medication information for the disease, International Classification of Diseases (ICD)-9 and -10 codes linked to Hospital Episode Statistics (refer to supplementary table S4 for the specific codes used) available at their initial visit, as in the previous studies of UK Biobank [11, 12]. We acknowledge that common diseases such as COPD and asthma were generally managed in primary care settings, and thus we included self-reported physician-made diagnoses in the disease ascertainment criteria in addition to Hospital Episode Statistics. The UK Biobank study protocol is available online [13]. The curated diagnoses were all known complications of AATD (supplementary table S4) [14–17]. ICD-10 codes in UK Biobank do not have subclassification of AATD by ICD-10 coding (E88.01), but do provide diagnosis of E88.0, which represents the combined diagnoses of plasma-protein metabolism disorders and may include diagnoses, in addition to AATD, such as plasminogen deficiency and bisalbuminaemia. To estimate the prevalence of AATD diagnosis, we identified individuals reporting a physician-made diagnosis of AATD and/or the use of medication for AATD; or having an ICD-9 diagnostic code for AATD. The supplementary material provides the detailed definition of symptoms and the spirometry quality control (supplementary table S5).

### 2.4.4 Statistical analysis

Regression models were fitted to assess the associations of *SERPINA1* genotypes and clinical outcomes compared to PI*MM genotype. All the models were adjusted for age, sex, genotyping array, assessment centre and the first five principal components in order to

account for population structure. We subsequently stratified the participants by smoking status. Within each genotype group, the decrease of $FEV_1$ by age was estimated by linear regression of $FEV_1$ by age, adjusted for the same covariates as above and thus these were not derived by the longitudinal data. Survival analysis was performed using univariate Cox proportional hazard model to estimate the hazard of death. Detailed methods of smoking status definition and survival analysis are presented in the supplementary material (supplementary table S6).

We estimated the national prevalence of PI*ZZ genotype status in the UK, assuming that the allele frequency rates were not different between individuals of European ancestry in UK Biobank and the UK citizens of European ancestry. This could be an underestimate of the PI*ZZ genotype frequency, given that UK Biobank is not a population-representative cohort, as it recruited only those aged >40 years and has some healthy volunteer bias. Next, we used data from the Office of National Statistics [18] to estimate the proportion of British citizens of European ancestry and estimated the number of British individuals carrying the PI*ZZ genotype.

### 2.4.5 Sensitivity analyses

We included those with E88.0 in ICD-10 codings for the diagnoses of AATD and recalculated the prevalence of the diagnosed AATD. As UK Biobank included pairs of relatives, we removed one randomly selected participant from each pair related to the third degree (kinship coefficient $\geqslant 0.0442$), leaving 449991 unrelated participants, to assess the inflation of association affected by familial effects. Multivariate Cox proportional hazard model adjusted for age was also applied for survival analysis.

### 2.4.6 Phenome-wide association study

Next, we explored associations of *SERPINA1* genotypes with other diseases using a PheWAS design. The detailed methods are described in the supplementary material (supplementary tables S7 and S8).

### 2.4.7 Polygenic risk score for $FEV_1$/FVC

The recent large scale genome-wide association study of spirometry data derived from external cohorts of European descent [19] enabled us to establish a PRS, the weighted sum of effect alleles of common variants that is associated with spirometry results. We calculated the $FEV_1$/FVC PRS of each individual and assessed the interactions between *SERPINA1* genotypes and this PRS. The detailed methodology is found in the supplementary material.

## 2.5. Results

### 2.5.1 Participant characteristics

We identified 458164 participants in UK Biobank of European descent who had a median age of 58 years (interquartile range (IQR) 50–63 years), and there were 61 (0.013%) people who were diagnosed as having AATD (table 1, supplementary figure S2). Among 140 participants with the PI*ZZ genotype, only nine (6.4%, 95% CI 3.4–11.7%) were diagnosed as AATD (table 2). Given that there are 65.6 million citizens of the UK [20], of whom 87% are estimated to be of European ancestry [18], we estimate that 17439 (95% CI 14671–20579) European individuals in the UK carry the PI*ZZ genotype.

Compared to those with PI*MM, participants with PI*ZZ had more respiratory symptoms (45% versus 25%), lower $FEV_1$/FVC (median 0.74 versus 0.77) and lower $FEV_1$ % predicted (median 86% versus 94%) (table 2). A total of 37 (37%) participants with PI*ZZ had

$FEV_1/FVC$ <0.7 (table 2). Among 17790 individuals with a diagnosis of COPD, 31 (0.17%) individuals had the PI*ZZ genotype, and they had more severe airway obstruction than PI*MM individuals. Lastly, only seven (23%, 95% CI 11–40%) PI*ZZ individuals with clinically detected COPD were diagnosed as having AATD (table 3), and among 1407 participants with a diagnosis of cirrhosis, three (0.21%) had the PI*ZZ genotype and none of them were diagnosed as AATD.

### 2.5.2 Association of PI*ZZ genotype with clinical outcomes

Those with PI*ZZ had a higher risk of COPD (OR 8.8, 95% CI 5.8–13.3; $p=1.1\times10^{-24}$), asthma (OR 2.0, 95% CI 1.4–3.0; $p=5.3\times10^{-4}$), bronchiectasis (OR 7.3, 95% CI 3.2–16.8; $p=2.4\times10^{-6}$) and pneumonia (OR 2.7, 95% CI 1.5–4.9; $p=1.2\times10^{-3}$) compared to PI*MM. Those with the PI*ZZ genotype had higher risk of COPD regardless of smoking status, but effect sizes were larger for smokers (OR 13.3, 95% CI 7.5–23.8versus OR 7.9, 95% CI 3.9–16.1). In never-smokers, the PI*ZZ genotype was not significantly associated with asthma or bronchiectasis (figure 1 and supplementary table S9). PI*ZZ was not independently associated with pneumonia when conditioned on the diagnosis of COPD (OR 1.5, 95% CI 0.8–2.8; p=0.21).

Among the extrapulmonary diseases we curated, PI*ZZ genotype was associated with diagnoses of cirrhosis (OR 7.8, 95% CI 2.5–24.6; p=0.004), hepatic carcinoma (OR 13.7, 95% CI 3.4–56.0; $p=2.7\times10^{-4}$) and panniculitis (OR 71.8, 95% CI 9.6–534.9; $p=3.1\times10^{-5}$) (supplementary table S9).

Individuals with PI*ZZ had more respiratory symptoms (OR 2.5, 95% CI 1.8–3.5; $p=6.5\times10^{-8}$) than PI*MM, such as wheeze (OR 2.1, 95% CI 1.5–3.0; $p=4.0\times10^{-5}$), shortness of breath (OR 3.3, 95% CI 1.9– 5.8; $p=3.0\times10^{-5}$), persistent cough (OR 4.2, 95% CI 2.2–7.8;

p=9.7×10$^{-6}$) and persistent sputum (OR 4.1, 95% CI 2.0–8.2; p=9.1×10$^{-5}$). For never-smokers, persistent cough was the only symptom associated with PI*ZZ (OR 3.3, 95% CI 1.3–8.9; p=0.016) (supplementary table S10). People with PI*ZZ genotype were more likely to have $FEV_1$/FVC <0.7 (OR 4.3, 95% CI 2.8–6.6; p=1.1×10$^{-11}$) and have $FEV_1$ <50% pred (OR 13.2, 95% CI 6.9–25.5; p=1.2×10$^{-14}$) (figure 1, supplementary table S10). Linear regression of $FEV_1$ by age estimated that the decrease of $FEV_1$ by age is 68.3 mL·year−1 (95% CI 47.1–89.7) in PI*ZZ participants compared to 35.6 mL·year−1 (95% CI 35.4–35.8) in PI*MM individuals (table 2, supplementary table S10). The difference of the decrease of $FEV_1$ by age between ever-smokers and never-smokers in PI*ZZ individuals was inconclusive because of the lack of statistical power (supplementary table S11). PI*ZZ genotype was associated with all-cause mortality compared to PI*MM genotype (hazard ratio 2.4, 95% CI 1.2–4.6; p=9.9×10$^{-3}$) during a median follow-up duration of 7.0 years (IQR 6.4–7.7 years) (figure 2, supplementary table S12). All results from sensitivity analyses are presented in the supplement (supplementary tables S12 and S13).

### 2.5.3 Phenome-wide association study

PI*ZZ genotype was associated with increased risk of other disorders of metabolism (including AATD), emphysema, obstructive chronic bronchitis and chronic airway obstruction (supplementary figure S3). In addition, PI*ZZ was associated with increased risks of dependency on a respirator or supplemental oxygen, empyema and pneumothorax, cachexia, polycythaemia, aneurysm and pancreatitis, all of which were statistically significant (p<6.1×10$^{-4}$) after Benjamini–Hochberg correction in the main analysis, conservatively assuming that all the phecodes tested were independent (supplementary figure S3 and table S14). The more detailed results of sensitivity analyses are in the supplementary material.

### 2.6.3 AATD-associated genotypes, other than PI*ZZ

Additionally, we analysed participants with PI*SZ, PI*MZ and PI*SS compared to PI*MM.

In brief, PI*SZ and PI*MZ genotypes were associated with a slight increase of $FEV_1$/FVC <0.7 (OR 1.3, 95% CI 1.0–1.6; p=0.022 and OR 1.1, 95% CI 1.0–1.1; p=0.032), but not associated with increased risk of clinically diagnosed COPD (figure 3, supplementary tables S9 and S10). Among heavy smokers (>20 pack-years), PI*SZ was associated with two-fold increased risk of $FEV_1$/FVC <0.7 (OR 2.0, 95% CI 1.3–3.1; $p=2.6\times10^{-3}$), whereas PI*MZ was associated with mildly increased risk of $FEV_1$/FVC <0.7 (OR 1.2, 95% CI 1.1–1.4; $p=4.5\times10^{-4}$) (supplementary table S10). PI*MZ was also associated with increased risk of cirrhosis (OR 1.5, 95% CI 1.2–1.8; p=0.002) (figure 3), hepatitis (OR 1.4, 95% CI 1.1–1.8; $p=4.6\times10^{-3}$) and granulomatosis with polyangiitis (OR 2.2, 95% CI 1.2–3.9; $p=9.9\times10^{-3}$) (supplementary table S9). All the other results are provided in the supplementary material (supplementary figures S4 and S5 and tables S9–S12, S15–S17).

### 2.6.4 Polygenic risk score for $FEV_1$/FVC

The square of the correlation coefficient (r2) between observed $FEV_1$/FVC and $FEV_1$/FVC predicted by the PRS was 3.5% (95% CI 3.4%–3.6%) in the total population (n=328638), which was higher than the correlation between $FEV_1$/FVC and smoking status (2.4%, 95% CI 2.3%–2.5%). The PRS was not associated with other nongenetic risk factors (supplementary table S18). We divided participants into quartiles according to their PRS (figure 4). Among PI*ZZ individuals, those with the lowest quartile of PRS, i.e. those at lowest polygenic risk (n=29), had higher $FEV_1$/FVC results compared to other PI*ZZ individuals (n=72) (median (IQR) 0.79 (0.67–0.85) versus 0.72 (0.66–0.77), p=0.019). Multivariate logistic regression indicated that 1-SD change of PRS and PI*ZZ are independently associated with $FEV_1$/FVC

<0.7 (OR 1.4, 95% CI 1.4–1.5; $p<2\times10^{-16}$ and OR 4.5, 95% CI 3.0–6.9; $p=2.3\times10^{-12}$, respectively) (supplementary table S19).

## 2.7 Discussion

Undertaking a large-scale assessment of the prevalence of *SERPINA1* genotypes, their associated odds of morbidity and mortality and the diagnostic rates of AATD in UK Biobank, we found that the vast majority of individuals with PI*ZZ were not diagnosed as having AATD. Yet, these individuals had substantially increased odds of respiratory symptoms, diseases and all-cause mortality. We estimated that ~17000 individuals in the UK carry the PI*ZZ genotype, which was similar to the estimates from the prior population-based neonatal screening studies [21, 22]. Nevertheless, this could be an underestimate given that UK Biobank recruited only those aged >40 years, and very ill individuals are unlikely to be able to take part. Thus, while the proportion of all British individuals who could be detected through genotyping efforts is small, the absolute number is not.

The impact of PI*ZZ genotype on health status is striking. PI*ZZ was associated with increased risk of COPD and pneumonia regardless of smoking status, yet the effect sizes for COPD were substantially larger among smokers. Furthermore, PI*ZZ genotype was associated with increased risk of asthma and bronchiectasis only among smokers. This suggests that smoking cessation has the potential to prevent those with PI*ZZ genotype from developing multiple respiratory diseases.

Almost half of those with the PI*ZZ genotype were symptomatic with severe airflow obstruction and increased risk of all-cause mortality. Linear regression of $FEV_1$ by age in PI*ZZ individuals estimated a larger age-dependent decrease of $FEV_1$ compared to PI*MM

individuals. In PheWAS, PI*ZZ was significantly associated with dependence on a respirator or supplemental oxygen, empyema and pneumothorax, cachexia and secondary polycythaemia, all of which could be sequelae of AATD. Extrapulmonary diseases that have previously been described as associated with PI*ZZ were also replicated in our study, such as cirrhosis, hepatic carcinoma, panniculitis, pancreatitis and aneurysm, pathogenesis of which is thought to be triggered by protease–antiprotease imbalance [23].

Even among subjects with COPD diagnosis, 77% of PI*ZZ individuals were not diagnosed as having AATD in this study. Previous surveys indicated that the mean delay between symptom onset and diagnosis among those actually diagnosed ranges from 5 to 8 years [6, 24], and the delay was associated with worse respiratory symptoms and accelerated emphysema progression [25]. Potential reasons for underdiagnosis include poor awareness of the disease, the unavailability of appropriate tests and/or treatments in specific regions, i.e. no availability of AAT replacement therapy in the UK [26, 27]. The current laboratory testing practice for AATD involves first quantifying plasma AAT levels together with measuring C-reactive protein, followed by protein phenotyping and/or Z and S genotyping [28, 29]. Since the genotype data is less affected by batch effects compared to measuring AAT, a protein known to increase in the context of inflammatory conditions [30], our results suggest that genotyping could be a step toward efficient identification of PI*ZZ carriers. In the current study, 80% of diagnosed AATD occurred in PI*MZ individuals. This could reflect either misdiagnoses or the impact of other disease predisposing mutations in the *SERPINA1* gene that were not detected with the genotyping array.

PI*ZZ individuals with the lowest quartile of the PRS had relatively higher $FEV_1$/FVC, possibly suggesting that polygenic factors affecting lung function partially explain variable penetrance of PI*ZZ genotype [3]. Genome-wide genotyping, which enables the calculation

of the PRS and the *SERPINA1* genotyping, could be alternative approach to the *SERPINA1*-targeted genotyping as a screening strategy for AATD, given its relatively low cost (USD 40 in a research context).

In addition, our study provides several insights of the effects of PI*SZ and PI*MZ genotype. Overall, while PI*SZ was associated with a two-fold increased risk of an $FEV_1/FVC < 0.7$ in heavy smokers, we demonstrated that PI*SZ and PI*MZ genotypes had modest effects on the risks of spirometry-defined obstructive lung impairment ($FEV_1/FVC < 0.7$) and severe airways obstruction ($FEV_1$ <50% pred) compared to the previous findings that PI*SZ had three-fold increased risk of COPD (95% CI 1.24–8.57) [5] and PI*MZ had five-fold increased risk of COPD (95% CI 1.27–21.15) [31]. However, these case– control studies described very large confidence intervals and the PI*MZ participants were recruited from index PI*MZ COPD patients [31], potentially biased by the other shared genetic factors associated with COPD. PI*MZ genotype, but not PI*SZ, was significantly associated with increased risk of cirrhosis and marginally increased risk of all-cause mortality in this study. The discordance between PI*SZ and PI*MZ genotype could be driven by the lack of statistical power in PI*SZ individuals, 20 times less than PI*MZ. PheWAS found that PI*MZ was associated with multiple diseases, namely increased risk of cholelithiasis and decreased risk of cardiovascular disease. There are several studies [32–35] which might support these hypotheses, although validation studies and functional investigations are necessary.

Most of the previous epidemiological studies of PI*ZZ individuals were case–control studies [36, 37] and the current study is the one of the largest studies to assess the effects of the *SERPINA1* genotype status to multiple health conditions in a single large population cohort. A prior family-based study included nonindex family members with undiagnosed PI*ZZ individuals who had more severe spirometry results (mean $FEV_1/FVC$ 0.61 and mean $FEV_1$

72.3% pred) [38] than those in UK Biobank (table 2), which could reflect the effects of other shared genetic factors. The main limitation of this study is that UK Biobank is not representative of general population as there is well-documented evidence of a "healthy volunteer' bias [39]. Therefore, we did not try to derive generalisable disease prevalence, but aimed to report the associations with PI*ZZ genotype and multiple health conditions. Another shortcoming is that the diagnosis of AATD was based on questionnaires and/or Hospital Episode Statistics, which rely on the diagnosis of each clinician and potentially harbour "clinical order" bias [40]. Nevertheless, the estimated prevalence of asthma (14%), COPD (3.9%) and bronchiectasis (0.69%) in PI*MM individuals were similar to the previous reports [41–43], which might support the validity of our approach of how to ascertain the disease status. PheWAS demonstrated that PI*ZZ was associated with increased risk of cystic fibrosis, which could represent misdiagnoses of bronchiectasis. PheWAS, which is based on ICD codings, can be underpowered, so that while no significant associations between PI*ZZ and liver diseases or asthma were observed, this does not preclude smaller effects. Last, there are no AAT measurements available in UK Biobank, so we could not test whether people with high-risk genotypes had low levels of plasma AAT. Although we did not test cost-effectiveness of the population-level screening of AATD, genome-wide genotyping may help the screening of individuals at risk, such as heavy smokers or with a family history of pulmonary disease, to identify those with undiagnosed AATD. As this is a genetic study with potential clinical implications, future effort is needed to address the issue of incidental findings, such as applying the American College of Medical Genetics and Genomics [44] recommendations as to how to report secondary findings.

In summary, we provide evidence that the vast majority of individuals with PI*ZZ are not diagnosed as having AATD, according to definitions available in UK Biobank. Yet these individuals have a profoundly increased burden of multiple symptoms and diseases and an

increased risk of all-cause mortality. Identification of these individuals could help to target smoking cessation programmes [45] and the ascertainment of family members, as well as disease-specific therapies [9]. Our data provide potential avenues to realise clinical benefits of emerging nationwide genomic efforts in the UK.

## 2.8 Figures

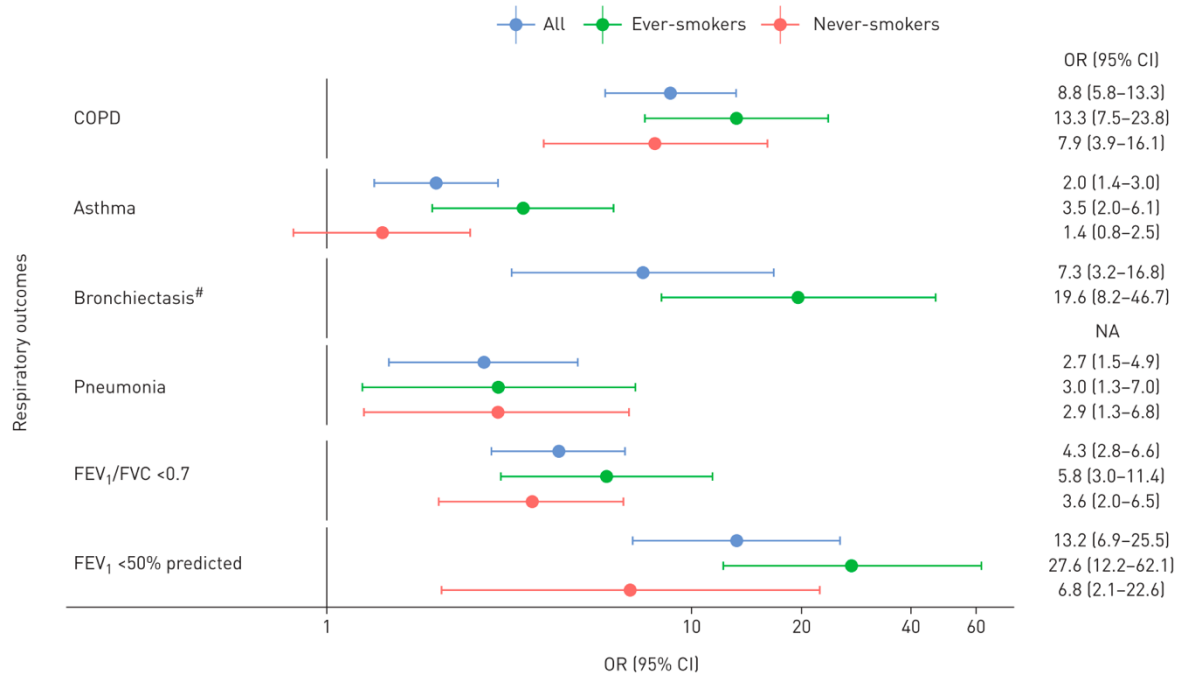


**Figure 1. Forest plot of associations between the PI*ZZ genotype and prevalent conditions stratified by smoking status.**

Odds ratios were calculated by logistic regression models compared to the PI*MM (wild-type) genotype adjusted for age, sex, genotyping array, assessment centre and the first five genetic principal components. $FEV_1$: forced expiratory volume in 1 s; FVC: forced vital capacity; NA: not applicable. #: no never- smokers have been diagnosed with bronchiectasis.

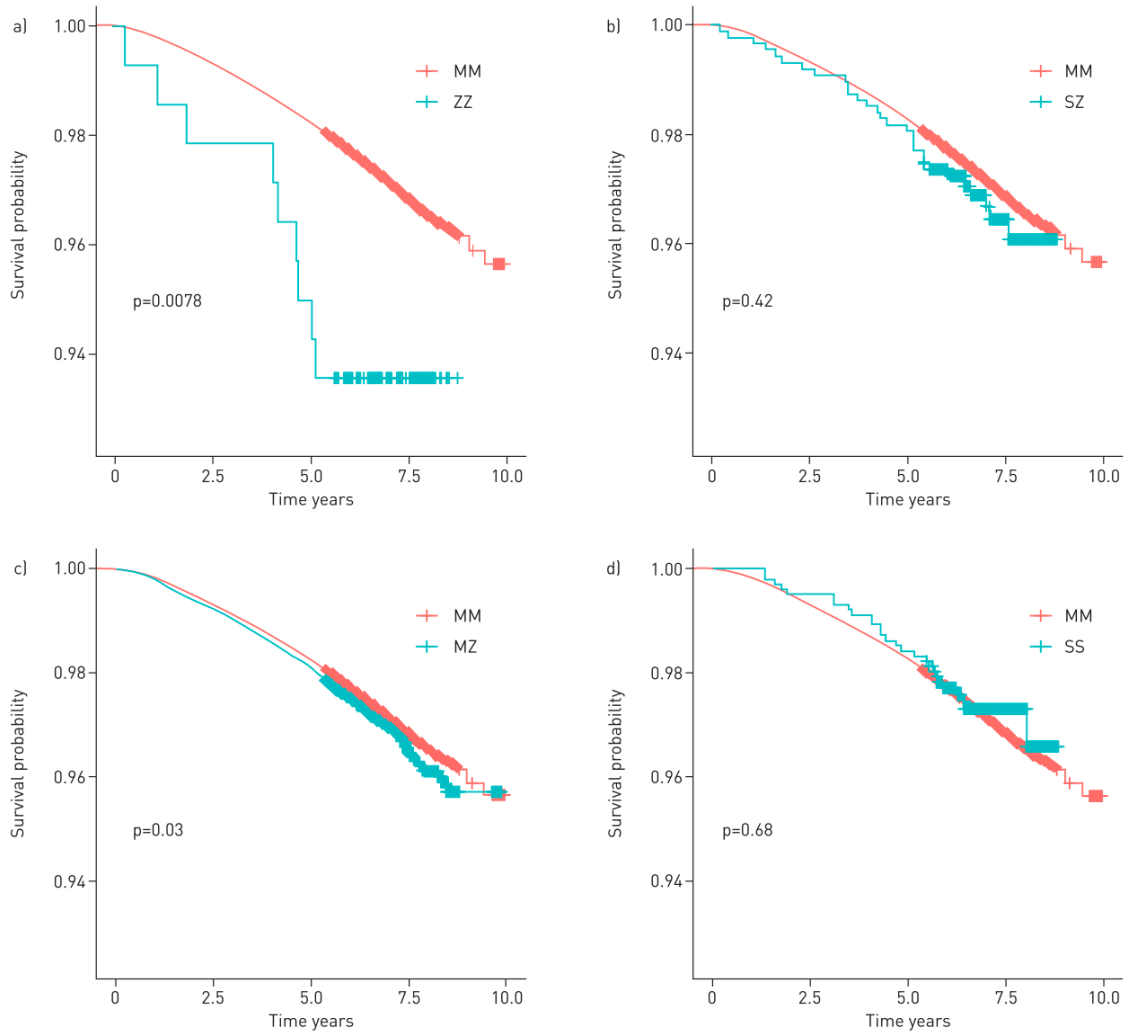


**Figure 2. Survival curves of all-cause mortality stratified by *SERPINA1* genotypes.**

a) PI*ZZ versus PI*MM genotypes; b) PI*SZ versus PI*MM genotypes; c) PI*MZ versus PI*MM genotypes; d) PI*SS versus PI*MM genotypes. All p-values were calculated by log-rank test.

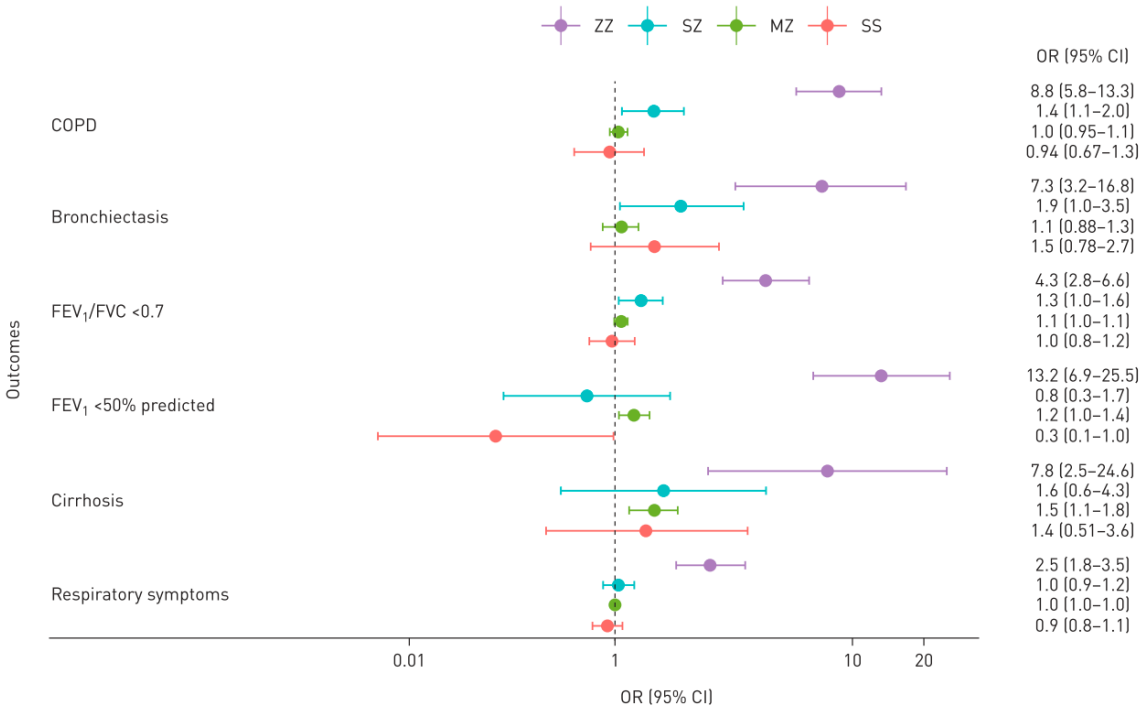


**Figure 3. Forest plot of associations between *SERPINA1* genotypes and common conditions.**

Odds ratios were calculated by logistic regression models compared to PI*MM (wild-type) genotype adjusted for age, sex, genotyping array, assessment centre and the first five genetic principal components. $FEV_1$: forced expiratory volume in 1 s; FVC: forced vital capacity.

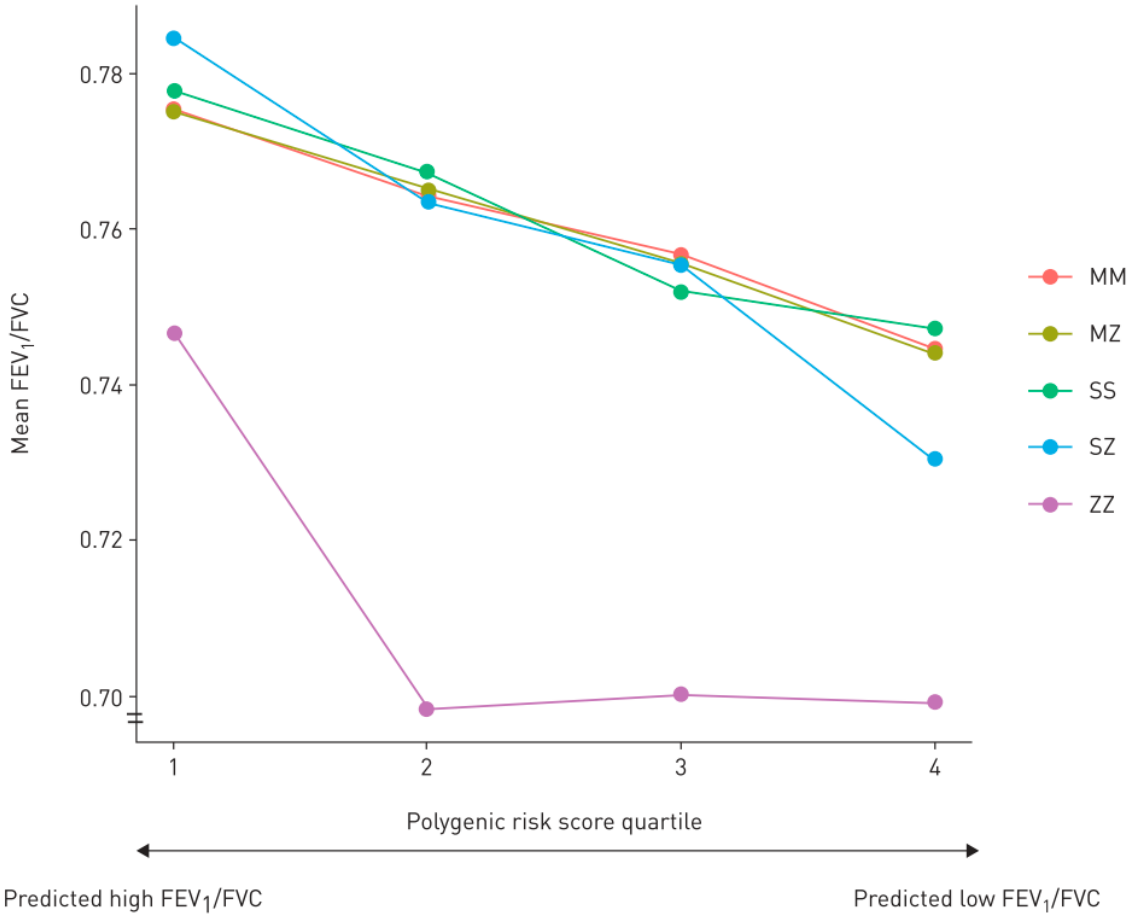


**Figure 4. Mean of observed forced expiratory volume in 1 s ($FEV_1$)/forced vital capacity (FVC) stratified by polygenic risk score quartile.**

Polygenic risk scores were calculated by LDpred using genome-wide association study summary statistics for $FEV_1$/FVC derived from the SpiroMeta consortium, which consists of individuals of European descent. Detailed methods are described in the supplementary material.

## 2.9 Tables

**Table 1. Participant characteristics stratified by *SERPINA1* genotypes.**

| | Total | Genotype | | | | | |
|---|---|---|---|---|---|---|---|
| | | MM | ZZ | SZ | MZ | SS | MS |
| **Subjects** | 458 164 (100) | 398 424 (87) | 140 (0.031) | 867 (0.19) | 16 983 (3.7) | 1013 (0.22) | 40 737 (8.9) |
| Unrelated individuals[#] | 449 991 | 391 334 | 138 | 851 | 16 707 | 993 | 39 968 |
| **Age years** | 58 (50–63) | 58 (50–63) | 56 (49–63) | 57 (50–63) | 58 (51–64) | 57 (50–63) | 58 (50–63) |
| **Male** | 209 694 (46) | 182 344 (46) | 73 (52) | 393 (45) | 7715 (45) | 469 (46) | 18 700 (46) |
| **Height cm** | 168.7±9.2 | 168.6±9.2 | 172.2±9.3 | 170.1±9.3 | 169.6±9.3 | 169.1±9.1 | 168.8±9.2 |
| Subjects with no height data | 1032 (0.23) | 898 (0.23) | 0 | 1 (0.12) | 36 (0.21) | 3 (0.30) | 94 (0.23) |
| **BMI kg·m$^{-2}$** | 27.4±4.8 | 27.4±4.8 | 26.7±4.7 | 27.0±4.6 | 27.3±4.7 | 27.3±4.6 | 27.4±4.8 |
| Subjects with no BMI data | 1522 (0.33) | 1321 (0.33) | 0 | 3 (0.35) | 50 (0.29) | 4 (0.39) | 144 (0.35) |
| **Smoking status** | 451 157 (98) | 392 313 (98) | 135 (96) | 856 (99) | 16 736 (99) | 996 (98) | 40 121 (98) |
| Current smokers[¶] | 47 711 (11) | 41 735 (11) | 7 (5.2) | 83 (9.7) | 1605 (9.6) | 106 (11) | 4175 (10) |
| Pack-years | 25.5 (14.7–37.8) | 25.3 (14.7–38.0) | 10.3 (7.8–14.9) | 25.4 (16.2–35.0) | 24.6 (14.3–37.4) | 30.0 (17.5–39.0) | 25.2 (14.6–37.5) |
| Subjects with pack-years data[¶] | 38 309 (80) | 33 512 (80) | 4 (57) | 70 (84) | 1298 (81) | 85 (80) | 3340 (80) |
| Past smokers[¶] | 158 852 (35) | 138 053 (35) | 45 (33) | 313 (37) | 5915 (35) | 334 (34) | 14 192 (35) |
| Pack-years | 17.0 (9.0–29.5) | 17.0 (9.0–29.5) | 15.9 (8.0–17.8) | 16.3 (9.0–26.0) | 17.0 (8.8–30.0) | 18.0 (10.0–31.5) | 17.5 (9.0–29.7) |
| Subjects with pack-years data[¶] | 103 195 (65) | 89 632 (65) | 27 (60) | 189 (60) | 3852 (65) | 221 (66) | 9274 (65) |
| Never-smokers[¶] | 244 594 (54) | 212 525 (54) | 83 (61) | 460 (54) | 9216 (55) | 556 (56) | 21 754 (54) |
| **Exposure to smoke or polluted air** | 120 423 (26) | 104 642 (26) | 36 (26) | 223 (26) | 4505 (27) | 256 (26) | 10 761 (26) |
| **AATD diagnosis** | 61 (0.013) | 4 (0.0010) | 9 (6.4) | 9 (1.0) | 36 (0.21) | 0 | 3 (0.0074) |

Data are presented as n (%), n, median (interquartile range) or mean±SD. BMI: body mass index; AATD: alpha-1 antitrypsin deficiency. #: numbers of individuals were calculated by removing related individuals with kinship coefficients ⩾0.044, which were used in sensitivity analyses; ¶: percentage was calculated among people with information available.

**Table 2. Clinical diagnoses and spirometry results of participants stratified by *SERPINA1* genotype.**

| | MM | ZZ | p-value# | SZ | p-value# | MZ | p-value# | SS | p-value# |
|---|---|---|---|---|---|---|---|---|---|
| **Subjects** | 398 424 | 140 | | 867 | | 16 983 | | 1013 | |
| **Respiratory symptoms** | 97 970 (25) | 63 (45) | $1.8\times10^{-7}$ | 219 (25) | 0.64 | 4150 (24) | 0.24 | 234 (23) | 0.29 |
| **AATD diagnosis** | 4 (0.0010) | 9 (6.4) | $4.5\times10^{-29}$ | 9 (1.0) | $7.3\times10^{-22}$ | 36 (0.21) | $7.8\times10^{-46}$ | 0 | 1 |
| **COPD diagnosis** | 15 502 (3.9) | 31 (22) | $2.9\times10^{-15}$ | 46 (5.3) | 0.034 | 676 (4.0) | 0.44 | 36 (3.6) | 0.68 |
| **Asthma diagnosis** | 54 205 (14) | 33 (24) | $1.3\times10^{-3}$ | 118 (14) | 1 | 2343 (14) | 0.48 | 127 (13) | 0.34 |
| **Bronchiectasis diagnosis** | 2767 (0.69) | 6 (4.3) | $4.8\times10^{-4}$ | 11 (1.2) | 0.06 | 125 (0.74) | 0.51 | 10 (0.99) | 0.25 |
| **PFT** | 285 824 (72) | 101 (72) | 1 | 613 (71) | 0.50 | 12 110 (71) | 0.22 | 728 (72) | 0.94 |
| $FEV_1$ L | 2.8 (2.3–3.3) | 2.8 (2.1–3.4) | 0.28 | 2.8 (2.4–3.4) | 0.015 | 2.8 (2.3–3.4) | $3.1\times10^{-7}$ | 2.8 (2.3–3.3) | 0.21 |
| $FEV_1$/FVC | 0.77 (0.73–0.80) | 0.74 (0.66–0.79) | $3.5\times10^{-4}$ | 0.77 (0.72–0.81) | 0.67 | 0.77 (0.73–0.80) | 0.63 | 0.77 (0.73–0.80) | 0.74 |
| $FEV_1$ % predicted | 94 (83–103) | 86 (75–103) | $2.3\times10^{-4}$ | 95 (85–104) | 0.28 | 94 (84–104) | 0.062 | 94 (84–104) | 0.28 |
| Decrease in $FEV_1$ by age¶ $mL{\cdot}year^{-1}$ | 35.6 (35.4–35.8) | 68.4 (47.1–89.7) | $2.0\times10^{-4}$ | 36.0 (35.5–37.8) | 0.74 | 36.6 (35.5–37.8) | 0.074 | 34.0 (29.4–38.5) | 0.51 |
| $FEV_1$/FVC <0.7+ | 40 351 (14) | 37 (37) | $1.5\times10^{-8}$ | 107 (17) | 0.02 | 1799 (15) | 0.023 | 99 (14) | 0.75 |

Data are presented as n (%) or median (interquartile range), unless otherwise stated. AATD: alpha-1 antitrypsin deficiency; PFT: pulmonary function testing; $FEV_1$: forced expiratory volume in 1 s; FVC: forced vital capacity. #: calculated by comparing to PI*MM genotype; ¶: estimated with linear regression by age (95% CI) and not derived from the longitudinal data; +: percentage calculated among subjects with spirometry information available.

**Table 3. Comparison of characteristics for PI*ZZ and PI*MM genotypes among individuals with COPD.**

| | ZZ | MM | p-value |
|---|---|---|---|
| **Subjects** | 31 (0.17) | 15502 (95) | |
| **Age years** | 56 (49–63) | 62 (57–66) | 0.47 |
| **Male** | 21 (68) | 8016 (52) | 0.1 |
| **Height cm** | 173.5±9.9 | 167.8±9.2 | $1.5\times10^{-3}$ |
| No height information | 0 | 63 (0.41) | 1 |
| **BMI kg·m$^{-2}$** | 25.8±4.9 | 28.4±5.7 | $1.9\times10^{-3}$ |
| No BMI information | 0 | 100 (0.65) | 1 |
| **Respiratory symptoms** | 28 (90) | 11470 (74) | 0.04 |
| **AATD diagnosis** | 7 (23) | 2 (0.013) | $2.3\times10^{-18}$ |
| **Exposure to smoke or polluted air** | 10 (32) | 4173 (27) | 0.54 |
| **Smoking status** | 30 (97) | 15159 (98) | 0.51 |
| Current smokers# | 2 (6.7) | 4328 (29) | $7.1\times10^{-3}$ |
| Ex-smokers# | 19 (63) | 7249 (48) | 0.10 |
| Never-smokers# | 9 (30) | 3582 (24) | $4.7\times10^{-3}$ |
| **PFT** | 14 (45) | 9427 (61) | 0.095 |
| $FEV_1$ L | 1.5 (1.7–2.7) | 2.2 (1.7–2.7) | 0.073 |
| $FEV_1$/FVC | 0.48 (0.42–0.65) | 0.70 (0.62–0.76) | $4.1\times10^{-5}$ |
| $FEV_1$ % predicted | 45 (36–79) | 76 (62–90) | $3.8\times10^{-3}$ |

Data are presented as n (%), median (interquartile range) or mean±SD, unless otherwise stated. n=17790. BMI: body mass index; AATD: alpha-1 antitrypsin deficiency; PFT: pulmonary function testing; $FEV_1$: forced expiratory volume in 1 s; FVC: forced vital capacity. #: percentage was calculated among subjects with information on smoking status.

## 2.10 List of references


1 Silverman EK, Sandhaus RA. Alpha1-antitrypsin deficiency. *N Engl J Med* 2009; 360: 2749–2757.

2 Stoller JK, Aboussouan LS. A review of α1-antitrypsin deficiency. *Am J Respir Crit Care Med* 2012; 185: 246–259.

3 Silverman EK, Pierce JA, Province MA, *et al*. Variability of pulmonary function in alpha-1-antitrypsin deficiency: clinical correlates. *Ann Intern Med* 1989; 111: 982–991.

4 DeMeo DL, Campbell EJ, Brantly ML, *et al*. Heritability of lung function in severe alpha-1 antitrypsin deficiency. *Hum Hered* 2009; 67: 38–45.

5 Dahl M, Hersh CP, Ly NP, *et al*. The protease inhibitor PI*S allele and COPD: a meta-analysis. *Eur Respir J* 2005; 26: 67–76.

6 Campos MA, Wanner A, Zhang G, *et al*. Trends in the diagnosis of symptomatic patients with α1- antitrypsin deficiency between 1968 and 2003. *Chest* 2005; 128: 1179–1186.

7 De Serres FJ. Worldwide racial and ethnic distribution of α1-antitrypsin deficiency: summary of an analysis of published genetic epidemiologic surveys. *Chest* 2002; 122: 1818–1829.

8 Luisetti M, Seersholm N. α1-Antitrypsin deficiency 1: epidemiology of α1-antitrypsin deficiency. *Thorax* 2004; 56: 164–169.

9 Chapman KR, Burdon JGW, Piitulainen E, *et al*. Intravenous augmentation treatment and lung density in severe α1 antitrypsin deficiency (RAPID): a randomised, double-blind, placebo-controlled trial. *Lancet* 2015; 386: 360–368.

10 Turro E, Astle W, Megy K. Whole-genome sequencing of rare disease patients in a national healthcare system. *Nature* 2020; 583: 96–102.

11 Said MA, Verweij N, van der Harst P. Associations of combined genetic and lifestyle risks with incident cardiovascular disease and diabetes in the UK Biobank study. *JAMA Cardiol* 2018; 3: 693–702.

12 Pilling LC, Tamosauskaite J, Jones G, *et al*. Common conditions associated with hereditary haemochromatosis genetic variants: cohort study in UK Biobank. *BMJ* 2019; 364: k5222.

13 UK Biobank. 2007. UK Biobank: Protocol for a large-scale prospective epidemiological resource. www.ukbiobank. ac.uk/wp-content/uploads/2011/11/UK-Biobank-Protocol.pdf.

14 Eriksson S, Carlson J, Velez R. Risk of cirrhosis and primary liver cancer in alpha1-antitrypsin deficiency. *N Engl J Med* 1986; 314: 736–739.

15 McBean J, Sable A, Maude J, *et al*. Alpha1-antitrypsin deficiency panniculitis. *Cutis* 2003; 71: 205–209.

16 Mahr AD, Edberg JC, Stone JH, *et al*. Alpha1-antitrypsin deficiency-related alleles Z and S and the risk of Wegener's granulomatosis. *Arthritis Rheum* 2010; 62: 3760–3767.

17 Sun Z, Yang P. Role of imbalance between neutrophil elastase and α1-antitrypsin in cancer development and progression. *Lancet Oncol* 2004; 5: 182–190.

18 Office for National Statistics (ONS). Key Statistics and Quick Statistics for Local Authorities in the United Kingdom. Newport, Office for National Statistics, 2013; pp. 1–27.

19 Shrine N, Guyatt AL, Erzurumluoglu AM, *et al*. New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. *Nat Genet* 2019; 51: 481–493.

20 Office for National Statistics (ONS). *National Population Projections: 2016-based Statistical Bulletin.* 2017. www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationprojections/bulletins/nationalpopu lationprojections/2016basedstatisticalbulletin

21 O'Brien ML, Buist NRM, Murphey WH. Neonatal screening for alpha1-antitrypsin deficiency. *J Pediatr* 1978; 92: 1006–1010.

22 Sveger T. Liver disease in alpha1-antitrypsin deficiency detected by screening of 200,000 infants. *N Engl J Med* 1976; 294: 1316–1321.

23 Wang Q, Du J, Yu P, *et al*. Hepatic steatosis depresses alpha-1-antitrypsin levels in human and rat acute pancreatitis. *Sci Rep* 2015; 5: 17833.

24 Stoller JK, Sandhaus RA, Turino G, *et al*. Delay in diagnosis of α1-antitrypsin deficiency: a continuing problem. *Chest* 2005; 128: 1989–1994.

25 Tejwani V, Nowacki AS, Fye E, *et al*. The impact of delayed diagnosis of alpha-1 antitrypsin deficiency: the association between diagnostic delay and worsened clinical status. *Respir Care* 2019; 64: 915–

26 Greulich T, Vogelmeier CF. Alpha-1-antitrypsin deficiency: increasing awareness and improving diagnosis. *Ther Adv Respir Dis* 2016; 10: 72–84.

27 Horváth I, Canotilho M, Chlumský J, *et al*. Diagnosis and management of α1-antitrypsin deficiency in Europe: an expert survey. *ERS Open Res* 2019; 5: 00171-2018.

28 Miravitlles M, Dirksen A, Ferrarotti I, *et al*. European Respiratory Society statement: Diagnosis and treatment of pulmonary disease in α1-antitrypsin deficiency. *Eur Respir J* 2017; 50: 1700610.

29 Franciosi AN, Carroll TP, McElvaney NG. Pitfalls and caveats in α1-antitrypsin deficiency testing: a guide for clinicians. *Lancet Respir Med* 2019; 7: 1059–1067.

30 Sanders CL, Ponte A, Kueppers F. The effects of inflammation on alpha 1 antitrypsin levels in a national screening cohort. *COPD* 2018; 15: 10–16.

31 Molloy K, Hersh CP, Morris VB, *et al*. Clarification of the risk of chronic obstructive pulmonary disease in α1-antitrypsin deficiency PiMZ heterozygotes. *Am J Respir Crit Care Med* 2014; 189: 419–427.

32 Fadini GP, Menegazzo L, Rigato M, *et al*. NETosis delays diabetic wound healing in mice and humans. *Diabetes* 2016; 65: 1061–1071.

33 Dijk W, Kersten S. Regulation of lipoprotein lipase by *Angptl4*. *Trends Endocrinol Metab* 2014; 25: 146–155.

34 Ferkingstad E, Oddsson A, Gretarsdottir S, *et al*. Genome-wide association meta-analysis yields 20 loci associated with gallstone disease. *Nat Commun* 2018; 9: 5101.

35 Fähndrich S, Biertz F, Karch A, *et al*. Cardiovascular risk in patients with alpha-1-antitrypsin deficiency. *Respir Res* 2017; 18: 171.

36 Piitulainen E, Eriksson S. Decline in $FEV_1$ related to smoking status in individuals with severe α1-antitrypsin deficiency (PiZZ). *Eur Respir J* 1999; 13: 247–251.

37 Tanash HA, Nilsson PM, Nilsson JÅ, *et al*. Clinical course and prognosis of never-smokers with severe alpha-1-antitrypsin deficiency (PiZZ). *Thorax* 2008; 63: 1091–1095.

38 DeMeo DL, Sandhaus RA, Barker AF, *et al*. Determinants of airflow obstruction in severe alpha-1-antitrypsin deficiency. *Thorax* 2007; 62: 805–812.

39 Fry A, Littlejohns TJ, Sudlow C, *et al*. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *Am J Epidemiol* 2017; 186: 1026–1034.

40 Bush WS, Oetjens MT, Crawford DC. Unravelling the human genome–phenome relationship using phenome-wide association studies. *Nat Rev Genet* 2016; 17: 129–145.

41 Health and Social Care Information Centre. *Health Survey for England 2018 Asthma.* 2019. healthsurvey.hscic.gov. uk/media/81643/HSE18-Asthma-rep.pdf Date last updated: December 3, 2019. Date last accessed: June 4, 2020.

42 Health and Safety Executive (HSE). 2019. *Work-related Chronic Obstructive Pulmonary Disease (COPD) statistics in Great Britain*, 2019. https://www.hse.gov.uk/statistics/causdis/copd.pdf.

43 Quint JK, Millett ERC, Joshi M, *et al*. Changes in the incidence, prevalence and mortality of bronchiectasis in the UK from 2004 to 2013: a population-based cohort study. *Eur Respir J* 2016; 47: 186–193.

44 Kalia SS, Adelman K, Bale SJ, *et al*. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genet Med* 2017; 19: 249–255.

45 Seersholm N, Kok-Jensen A. Survival in relation to lung function and smoking cessation in patients with severe hereditary alpha 1-antitrypsin deficiency. *Am J Respir Crit Care Med* 1995; 151: 369–373.

## 2.11 Supplemental data

Supplementary Methods, Tables and Figures can be downloaded from the open access publication Nakanishi et al. in *Eur Respir J* available here: https://erj.ersjournals.com/content/56/6/2001441#sec-16

**Connecting Text: Bridge Between Chapter 2 and Chapter 3**

In the previous Chapter, we undertook a large-scale assessment of the prevalence of *SERPINA1* genotypes in UK Biobank and provided evidence that the vast majority of individuals with PI*ZZ are not diagnosed as having AATD. Yet these individuals have a profoundly increased burden of multiple symptoms and diseases and an increased risk of all-cause mortality. Identification of these individuals through genetic testing could help to target smoking cessation programs and the ascertainment of family members, as well as disease-specific therapies.

In the next Chapter, we hypothesized that genetic testing may be also helpful in the clinical management of COVID-19, the emerging respiratory illness caused by SARS-CoV-2 infection. The variability in clinical outcomes of COVID-19 causes difficulties in clinical management when estimating who is at risk of severe disease and may develop a need for intensive care. Furthermore, recent guidelines suggest risk stratification should be considered when deciding upon prophylactic treatment. In Chapter 3, we combined individual-level data from 13,888 COVID-19 patients from 17 cohorts in 9 countries to assess the association of the major common COVID-19 genetic risk factor with mortality, COVID-19-related complications, and laboratory values. By leveraging the large-scale aggregation of studies of heterogeneous design, we assessed to what extent this genetic information could predict COVID-19 severity, compared to the other non-genetic risk factors, such as BMI or age.

# Chapter 3: Age-dependent impact of the major common genetic risk factor for COVID-19 on severity and mortality

## 3.1 Title page

**Age-dependent impact of the major common genetic risk factor for COVID-19 on severity and mortality**

Tomoko Nakanishi,[1,2,3,4,5], Sara Pigazzini,[1,6] Frauke Degenhardt,[7] Mattia Cordioli,[1] Guillaume Butler-Laporte,[3,8] Douglas Maya-Miles,[9,10] Luis Bujanda,[11] Youssef Bouysran,[12] Mari E.K. Niemi,[1] Adriana Palom,[13,14,15] David Ellinghaus,[7,16] Atlas Khan,[17] Manuel Martínez-Bueno,[18] Selina Rolker,[19] Sara Amitrano,[20] Luisa Roade Tato,[10,13,14] Francesca Fava,[20,21,22] FinnGen,[23] The COVID-19 Host Genetics Initiative (HGI),[24] Christoph D. Spinner,[25] Daniele Prati,[26] David Bernardo,[10,27] Federico Garcia,[28,29] Gilles Darcis,[30,31] Israel Fernández-Cadenas,[32] Jan Cato Holter,[33,34] Jesus M. Banales,[11,35] Robert Frithiof,[36] Krzysztof Kiryluk,[17] Stefano Duga,[37,38] Rosanna Asselta,[37,38] Alexandre C. Pereira,[39] Manuel Romero-Gómez,[9,10] Beatriz Nafría-Jiménez,[40] Johannes R. Hov,[33,41,42] Isabelle Migeotte,[12,43] Alessandra Renieri,[20,21,22] Anna M. Planas,[44,45] Kerstin U. Ludwig,[19] Maria Buti,[10,13,14] Souad Rahmouni,[29] Marta E. Alarcón-Riquelme,[18,46] Eva C. Schulte,[47,48,49] Andre Franke,[7,50] Tom H. Karlsen,[34,41,42] Luca Valenti,[51,52] Hugo Zeberg,[53,54] J. Brent Richards,[2,3,9,55#] and Andrea Ganna[1,56#]

**[#]JBR and AG contributed equally to this work.**

**Affiliations:** [1]Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland. [2]Department of Human Genetics and [3]Centre for Clinical Epidemiology, Department of Medicine, Lady Davis Institute, Jewish General Hospital, McGill University, Montréal, Quebec, Canada. [4]Kyoto-McGill International Collaborative School in Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan. [5]Japan Society for the Promotion of Science, Tokyo, Japan. [6]University of Milano-Bicocca, Milano, Italy. [7]Institute of Clinical Molecular Biology, Christian-Albrechts-University, Kiel, Germany. [8]Department

of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Quebec, Canada. [9]Digestive Diseases Unit, Virgen del Rocio University Hospital, Institute of Biomedicine of Seville, University of Seville, Seville, Spain. [10]Centro de Investigación Biomédica en Red en Enfermedades Hepáticas y Digestivas (CIBERehd), Instituto de Salud Carlos III (ISCIII), Madrid, Spain. [11]Department of Liver and Gastrointestinal Diseases, Biodonostia Health Research Institute, Donostia University Hospital, University of the Basque Country (UPV/EHU), CIBERehd, Ikerbasque, San Sebastian, Spain. [12]Centre de Génétique Humaine, Hôpital Erasme, Université Libre de Bruxelles (ULB), Brussels, Belgium. [13]Liver Unit, Department of Internal Medicine, Hospital Universitari Vall d'Hebron, Vall d'Hebron Barcelona Hospital Campus, Barcelona, Spain. [14]Departament de Medicina: Bellaterra, Universitat Autònoma de Barcelona, Barcelona, Spain. [15]Liver Diseases, Vall d'Hebron Institut de Recerca (VHIR), Barcelona, Spain. [16]Novo Nordisk Foundation Center for Protein Research, Disease Systems Biology, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [17]Division of Nephrology, Department of Medicine, Vagelos College of Physicians and Surgeons, Columbia University, New York, New York, USA. [18]GENYO, Centre for Genomics and Oncological Research: Pfizer/University of Granada/Andalusian Regional Government, Granada, Spain. [19]Institute of Human Genetics, University Hospital Bonn, Medical Faculty University of Bonn, Bonn, Germany. [20]Genetica Medica, Azienda Ospedaliero-Universitaria Senese, Siena, Italy. [21]Medical Genetics and [22]Med Biotech Hub and Competence Center, Department of Medical Biotechnologies, University of Siena, Italy. [23]FinnGen is detailed in Supplemental Acknowledgments. [24]The COVID-19 HGI is detailed in Supplemental Acknowledgments. [25]Technical University of Munich, School of Medicine, University Hospital Rechts der Isar, Department of Internal Medicine II, Munich, Germany. [26]Department of Transfusion Medicine and Hematology, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Università degli Studi di Milano, Milano, Italy. [27]Mucosal Immunology Lab, Unit of

Excellence Institute of Biomedicine and Molecular Genetics (IBGM), University of Valladolid-CSIC, Valladolid, Spain. [28]Hospital Universitario Clinico San Cecilio, Granada, Spain. [29]Instituto de Investigación Ibs.Granada, Granada, Spain. [30]University of Liege, GIGA-Insitute, Liege, Belgium. [31]Liege University Hospital (CHU of Liege), Liege, Belgium. [32] Stroke Pharmacogenomics and Genetics Group, Biomedical Research Institute Sant Pau (IIB Sant Pau), Barcelona, Spain. [33]Department of Microbiology, Oslo University Hospital, Oslo, Norway. [34] Institute of Clinical Medicine, University of Oslo, Oslo, Norway. [35]Department of Biochemistry and Genetics, School of Sciences, University of Navarra, Pamplona, Spain. [36] Department of Surgical Sciences, Anaesthesiology and Intensive Care Medicine, Uppsala University, Uppsala, Sweden. [37]Department of Biomedical Sciences, Humanitas University, Pieve Emanuele, Milan, Italy. [38]IRCCS Humanitas Clinical and Research Hospital, Rozzano, Milan, Italy. [39]Heart Institute (InCor)/University São Paulo Medical School, São Paulo, Brazil. [40]Osakidetza Basque Health Service, Donostialdea Integrated Health Organisation, Clinical Biochemistry Department, Sebastian, Spain. [41]Norwegian PSC Research Center and Section of Gastroenterology, Department of Transplantation Medicine and [42]Research Institute of Internal Medicine, Oslo University Hospital, Oslo, Norway. [43]Fonds de la Recherche Scientifique (FNRS), Brussels, Belgium. [44]Institute for Biomedical Research of Barcelona (IIBB), National Spanish Research Council (CSIC), Barcelona, Spain. [45]Institut d'Investigacions Biomediques August Pi i Sunyer (IDIBAPS), Barcelona, Spain. [46]Institute for Environmental Medicine, Karolinska Institutet, Solna, Sweden. [47]Institute of Virology, Technical University of Munich/Helmholtz Zentrum München, Munich, Germany. [48]Institute of Psychiatric Phenomics and Genomics and [49]Department of Psychiatry, University Hospital, LMU Munich University, Munich, Germany. [50]University Hospital Schleswig-Holstein, Campus Kiel, Kiel, Germany. [51]Department of Pathophysiology and Transplantation, Università degli Studi di Milano, Milano, Italy. [52]Department of Transfusion

Medicine and Hematology, Precision Medicine, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milano, Italy. [53]Department of Neuroscience, Karolinska Institutet, Sweden. [54]Max-Planck Institute for Evolutionary Anthropology, Leipzig, Germany. [55]Department of Twin Research, King's College London, London, United Kingdom. [56]Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA

**Published in:** *J Clin Invest*. 2021;131(23):e152386. doi: 10.1172/JCI152386

## 3.2 Abstract

**BACKGROUND.** There is considerable variability in COVID-19 outcomes among younger adults, and some of this variation may be due to genetic predisposition.

**METHODS.** We combined individual level data from 13,888 COVID-19 patients (n = 7185 hospitalized) from 17 cohorts in 9 countries to assess the association of the major common COVID-19 genetic risk factor (chromosome 3 locus tagged by rs10490770) with mortality, COVID-19-related complications, and laboratory values. We next performed metaanalyses using FinnGen and the Columbia University COVID-19 Biobank.

**RESULTS.** We found that rs10490770 risk allele carriers experienced an increased risk of all-cause mortality (HR, 1.4; 95% CI, 1.2–1.7). Risk allele carriers had increased odds of several COVID-19 complications: severe respiratory failure (OR, 2.1; 95% CI, 1.6–2.6), venous thromboembolism (OR, 1.7; 95% CI, 1.2–2.4), and hepatic injury (OR, 1.5; 95% CI, 1.2–2.0). Risk allele carriers age 60 years and younger had higher odds of death or severe respiratory failure (OR, 2.7; 95% CI, 1.8–3.9) compared with those of more than 60 years (OR, 1.5; 95% CI, 1.2–1.8; interaction, P = 0.038). Among individuals 60 years and younger who died or experienced severe respiratory failure, 32.3% were risk-variant carriers compared with 13.9% of those not experiencing these outcomes. This risk variant improved the prediction of death or severe respiratory failure similarly to, or better than, most established clinical risk factors.

**CONCLUSIONS.** The major common COVID-19 genetic risk factor is associated with increased risks of morbidity and mortality, which are more pronounced among individuals 60 years or younger. The effect was similar in magnitude and more common than most established clinical risk factors, suggesting potential implications for future clinical risk management.

## 3.3 Introduction

The COVID-19 pandemic has led to the deaths of millions of individuals and the largest economic contraction since the Great Depression (1). The clinical outcomes of COVID-19 are remarkably variable, such that some individuals remain asymptomatic (2), while others develop severe COVID-19 with systemic inflammation, respiratory failure, or death. This variability in outcome creates difficulties in clinical management when estimating who is at risk of severe disease and may develop a need for intensive care. Furthermore, recent guidelines suggest risk stratification should be considered when deciding upon prophylactic treatment (3–5).

Some of this variation in COVID-19 behavior has been attributed to risk factors such as age, sex (6), comorbidities (7), socioeconomic factors (8), and genetic variants in the SARS-CoV-2 genome (9). While the main risk factor for severe outcomes is age, the impact of which increases exponentially after age 60 (7), some younger individuals experience severe COVID-19 outcomes and death. The early onset of several common diseases, such as breast cancers, myocardial infarction, and Alzheimer's disease, is disproportionally influenced by human genetic factors (10–13), and this may also be the case for COVID-19. Several GWAS have identified multiple loci in the human genome associated with severity of COVID-19 (14–17). Among GWAS findings, a genetic risk locus on chromosome 3 is the strongest and most consistent signal (16). This genetic risk locus harbors a cluster of genes on chromosome 3; however, the true causal variant is still unknown. The fact that the risk allele sits on a long haplotype inherited from Neanderthals (18) makes the identification of the causal allele and the gene or genes involved challenging. The SNP rs10490770 serves as a marker for this genetic risk factor (as well as other SNPs on the same haplotype; ref. [19]), and approximately 15% of individuals of European ancestry carry the C risk allele (19). However, the clinical relevance of this locus and its potential age-dependent impact are unknown.

We therefore assembled individual-level COVID-19 clinical and human genomic data in a large international consortium of 17 cohorts in 9 countries (Belgium, Brazil, Canada, Germany, Italy, Norway, Spain, Sweden, and the United Kingdom) to assess the relationship between the chromosome 3 SNP rs10490770 and COVID-19 severity, complications, and mortality, focusing on age-dependent effects. Finally, in order to assess the relative importance of this locus, we compared its ability to predict COVID-19 outcomes to that of a polygenic risk score (PRS), which aggregates information from common genetic variants across the genome, and other established clinical risk factors.

## 3.4 Results

### 3.4.1 Study participants

We collected and harmonized individual-level clinical and genomic data from 13,888 COVID-19 patients diagnosed with COVID-19 from February 5, 2020, to February 7, 2021. Table 1 illustrates the participants' demographic and clinical characteristics. By genetically inferring the ancestry using 1000G genetic superpopulations (20) as a reference, the majority of participants were of European descent (12,091; 87.1%). However, considerable numbers of individuals who were not of European were also included in metaanalyses: 389 (2.8%) were of South Asian ancestry, and 602 (4.3%) were of admixed American ancestry. Of these patients, 7185 were hospitalized, among whom 1695 (24.3%) were admitted to the intensive care unit (ICU); 1264 (10.0%) died following COVID-19 diagnosis, and 1704 (14.6%) met the criteria for severe respiratory failure (noninvasive ventilation, high-flow oxygen therapy, or intubation); their mean age was 62.9 years, and 31.2% were females. Clinical information was obtained with different degrees of completeness across studies. A detailed description of study-specific demographics, clinical characteristics, and their missingness rates is provided

in Supplemental Figure 1 and Supplemental Table 1 (supplemental material available online with this article; https://doi.org/10.1172/JCI152386DS1).

### 3.4.2 Chromosome 3 genetic risk and a PRS.

In order to tag the chromosome 3 locus, we selected the SNP rs10490770, which was most significantly associated with hospitalization in the COVID-19 GWAS from the COVID-19 HGI, since this is the largest GWAS metaanalysis of COVID-19 severity (16) (cases/controls = 12,888/1,295,966). We then compared the predictive performance of rs10490770 and a PRS. Using the COVID-19 HGI GWAS release 6 (https://www.covid19hg.org/results/r6/), we first metaanalyzed GWAS results from cohorts that were not included in our study (Supplemental Table 2) and calculated PRSs using a pruning and thresholding method. A PRS of $P = 5 \times 10^{-4}$ and $r = 0.7$ had the maximum accuracy in prediction for death or severe respiratory failure and was more significantly associated with death or severe respiratory failure than the chromosome 3 SNP only (OR, 1.7 vs. 1.2 per 1 SD increase in PRS and rs10490770, respectively; Supplemental Tables 3 and 4). Nevertheless, we focused on exploring the clinical implications of rs10490770, given that a single variant can be more easily tested in a clinical context, requires fewer computational resources than a PRS, and is less influenced by limitations, such as the poor transferability of PRSs across different ancestry groups.

### 3.4.3 Risk allele frequency.

We applied a dominant model by grouping participants into 2 groups according to their genotype at rs10490770; C is the allele associated with COVID-19 severity. Those with TC genotype or CC genotype were labeled as carriers, and those with TT genotype were labeled as noncarriers. According to the population frequencies in gnomAD (19), we estimate that 14.4% of individuals of European descent carry at least 1 rs10490770 C allele, as well as

9.5% of admixed Americans, 2.4% of Africans, 47.1% of South Asians, and 0.4% of East Asians. The carrier frequency was 16.2% among individuals of European descent in our cohort.

### 3.4.4 Association with mortality.

We first estimated the HR for all-cause mortality and COVID-19–related death. All analyses were performed separately for each ancestry group. Because the sample size in non-Europeans was limited, we reported the results from individuals of European descent as the main analyses, but the results from non-European ancestry individuals are presented in Supplemental Figures 4–7. All analyses were based on mixed-effects model adjusted for age, sex, and the first 5 genetic principal components (PCs) as fixed effects. Study groups were also included as random effects to account for the study variability.

Risk allele carriers at rs10490770 had a higher HR for all-cause mortality compared with noncarriers (HR, 1.4; 95% CI, 1.2– 1.7, P = $4.5 \times 10^{-5}$, dead/alive = 870/8829) over a median follow-up duration of 43 days (IQR, 17.5–69 days; Figure 1A). A competing risk model to estimate the HR for COVID-19–related death while accounting for non-COVID-19–related deaths estimated a similar HR for COVID-19–related mortality (HR, 1.6; 95% CI, 1.3–1.8, P = $4.5 \times 10^{-7}$, dead/alive = 750/8829; Figure 1B). The association with mortality was reduced, but still significant, when the analysis was restricted to hospitalized individuals (HR for all-cause mortality, 1.2; 95% CI, 1.0–1.4, P = 0.03, dead/alive = 870/3206, and HR for COVID-19 related mortality, 1.3; 95% CI, 1.1–1.6, P = $1.1 \times 10^{-3}$, dead/alive = 750/3206), indicating that the effect of rs10490770 on mortality was not simply explained by the higher hospitalization rate among the carriers.

### 3.4.5 Associations with COVID-19 severity.

We next examined the effect of risk allele carrier status at rs10490770 for COVID-19 severity. We confirmed that risk allele carrier status at rs10490770 was significantly associated with hospitalization (OR, 1.5; 95% CI, 1.3–1.7, P = $1.2 \times 10^{-9}$, cases/controls = 6054/6004). A stronger effect was observed for ICU admission (OR, 2.5; 95% CI, 1.9–3.2, P = $1.6 \times 10^{-12}$, cases/controls = 1234/6004) and death or severe respiratory failure (OR, 1.7; 95% CI, 1.5–2.1, P = $9.0 \times 10^{-10}$, cases/controls = 2005/7047; Figure 2 and Supplemental Table 5). Restricting analyses to hospitalized individuals, we observed consistent results, some of which were with diminished effect sizes (Figure 2 and Supplemental Table 5). For instance, a significant reduction in effect size was observed in OR for ICU admission (OR, 1.6; 95% CI, 1.3–1.8, P = $3.5 \times 10^{-8}$, cases/controls = 1234/4820).

We next explored the association of the rs10490770 risk allele with laboratory values that are known to be associated with the severity of COVID-19 (21–25). rs10490770 risk allele carrier status was associated with the worst value for each of these laboratory values at hospital (e.g., lactate dehydrogenase: 0.23 SD increase, P = $3.5 \times 10^{-7}$, D-dimer: 0.14 SD increase, P = $2.1 \times 10^{-3}$; IL-6: 0.16 SD increase, P = $8.7 \times 10^{-3}$; Supplemental Table 6 and Supplemental Figures 2 and 3).

### 3.4.6 Associations with COVID-19 complications.

Risk allele carrier status at rs10490770 was associated with multiple COVID-19– related severe complications (Figure 2). These included severe respiratory failure (OR, 2.1; 95% CI, 1.6–2.6, P = $2.3 \times 10^{-10}$, cases/controls = 1284/7047), venous thromboembolism (VTE) (OR, 1.7; 95% CI, 1.2–2.4, P = $1.1 \times 10^{-3}$, cases/controls = 208/8,936), and hepatic injury (OR, 1.5; 95% CI, 1.2-2.0, P = $1.4 \times 10^{-3}$, cases/controls = 352/9541). No significant effect was observed for cardiovascular complications (OR, 1.2; 95% CI, 1.0–1.5, P = 0.10,

cases/controls = 854/8890), although this might be due to lack of statistical power to detect such effects. Similar results were observed when restricting the analyses to hospitalized patients (Figure 2 and Supplemental Table 5).

### 3.4.7 Age-dependent associations with COVID-19 severity.

We explored how the effects of rs10490770 risk allele carrier status on severe COVID-19 outcomes in individuals of European descent varied by age. Among severe patients who died or had severe respiratory failure, rs10490770 risk allele carriers were on average 2.3 (95% CI, 1.1–3.5) years younger than noncarriers ($P = 1.6 \times 10^{-4}$, n = 2005; Figure 3A and Supplemental Table 5). Stratifying by age, we found that among those who were 60 years or younger, risk allele carrier status had markedly increased odds of death or severe respiratory failure (OR, 2.7 95% CI, 1.8–3.9), whereas risk allele carrier status had more modest effects among those older than 60 years with an OR of 1.5 (95% CI, 1.2–1.9, P value interaction = 0.038; Figure 3B and Supplemental Tables 5 and 7). Among all participants 60 years or younger who died or experienced a severe respiratory COVID-19 outcome, we found that 32.3% (95% CI, 28.3%–36.7%) were rs10490770 risk variant carriers, compared with 13.9% (95% CI, 12.6%–15.2%) of those who did not experience severe disease (Table 2). When considering other severity phenotypes, such as hospitalization and ICU admission, we observed that risk allele carriers tended to be younger than noncarriers. However, we did not detect a different effect in the association between rs10490770 risk allele carriers and these additional severity phenotypes among those who were 60 or younger versus more than 60 years old. This could be attributed to the heterogeneity of the criteria of hospitalization or ICU admission or case-control imbalance in some participating studies.

### 3.4.8 Associations with COVID-19 severity stratified by established clinical risk factors.

We studied how the effects of rs10490770 risk allele carrier status on COVID-19 severity varied by other established clinical risk factors. Among individuals with no risk factors (BMI ≥ 30, smoking, cancer, chronic kidney disease, chronic obstructive pulmonary disease (COPD), heart failure, transplantation, and diabetes mellitus [DM]) prior to COVID-19, risk allele carriers had an OR of 1.8 for death or severe respiratory failure (95% CI, 1.0–3.4), whereas risk allele carrier status had more modest effects among those with 1 risk factor (OR, 1.6; 95% CI, 1.1–2.5) and more than 1 risk factor (OR, 1.4; 95% CI, 1.0–1.8) (P value for interaction = 0.091; Figure 3B and Supplemental Table 8).

### 3.4.9 Risk prediction compared with established clinical risk factors.

We compared the risk discrimination conferred by the rs10490770 risk allele on COVID-19 severity with that observed for other established COVID-19 risk factors. To do so, we used multivariable regression in 7983 individuals of European ancestry, with complete ascertainment of clinical risk factors. rs10490770 risk allele carrier status was independent of other risk factors (Figure 4A and Supplemental Table 9) when examining the association with death or severe respiratory failure (OR, 2.0; 95% CI, 1.7–2.4, $P = 1.7 \times 10^{-13}$; frequency of risk allele carriers, 14.7%, cases/controls = 898/6454). The effect sizes were comparable, or larger, than those of other known risk factors such as DM (OR, 2.0; 95% CI, 1.7–2.4, $P = 1.0 \times 10^{-12}$, frequency of DM, 12.5%). Stronger effects were observed among individuals 60 years or younger (risk allele carrier status: OR, 3.5; 95% CI, 2.3–5.3, $P = 1.4 \times 10^{-9}$ ; frequency of risk allele carriers, 14.5%; cases/controls = 151/2348) relative to DM (OR, 2.7; 95% CI, 1.6–4.5, $P = 4.4 \times 10^{-4}$; frequency of DM, 5.7%; Figure 4A and Supplemental Table 9).

Consistent with the results from multivariable regression, adding the rs10490770 genotype to nongenetic risk factors modestly improved discrimination for death or severe respiratory failure among patients 60 years or younger (AUC: 0.82 vs. 0.84, P = 0.021, and net reclassification improvement [NRI], 0.41, $P = 7.7 \times 10^{-8}$; Table 3), and the performance of risk discrimination was similar to, or better than, that of most of established risk factors included in the study (Figure 4B and Supplemental Table 10).

### 3.4.10 Metaanalyses.

We next metaanalyzed the European ancestry results presented above with those of non-European ancestry participants and 2 external cohorts. We confirmed similar effects in the associations with mortality (Supplemental Figure 4), COVID-19 severity (Supplemental Figure 5), COVID-19 complications (Supplemental Figure 6), and age-dependent effects (Supplemental Figure 7). Given the small sample size of non-European participants, we lacked sufficient statistical power to investigate whether the association between rs10490770 risk allele carriers and COVID-19 outcomes was different when comparing individuals of non-European and European ancestry. Sensitivity analysis. Finally, we performed several sensitivity analyses to evaluate the robustness of our results. First, we removed the study variables from the covariates (Supplemental Tables 11 and 12). Second, we included participating studies themselves either as fixed or random effects (Supplemental Tables 11 and 12). Third, we restricted the analyses to individuals of European descent from UK Biobank (UKB), a cohort that was not developed to study COVID-19 and thus is less prone to selection bias. These UKB analyses generated similar results (Supplemental Table 13). Fourth, we explored different cutoffs for age-stratified analyses (Supplemental Table 14). Finally, we excluded related individuals (Supplemental Table 15). All sensitivity analyses were consistent with the results from the main analyses.

### 3.5. Discussion

Combining individual-level clinical and genomic data from 13,888 individuals ascertained for COVID-19 outcomes from 17 cohorts in 9 countries, we found that the major genetic risk factor for severe COVID-19 on chromosome 3 was strongly associated with COVID-19–related mortality and clinical complications, such as respiratory failure and VTE.

The risk allele is common. We estimated that 14.4% of individuals of European ancestry are risk allele carriers at rs10490770. Further, 9.5% of admixed Americans, 2.4% of Africans, 47.1% of South Asians, and 0.4% of East Asians are risk allele carriers (20). Consequently, a large proportion of humans carry this risk factor.

The effect of carrying the risk allele on COVID-19 severity was stronger in younger individuals. First, among those 60 years or younger, the odds of death or severe respiratory failure increased 2.7-fold for risk allele carriers. We found that 32% of individuals 60 years or younger who died or experienced severe respiratory failure were risk allele carriers compared with 14% of individuals not requiring supplemental oxygen. Second, among individuals who died or experienced severe respiratory failure, risk allele carriers were on average 2.3 years younger than noncarriers. Finally, the risk discrimination for death and severe respiratory COVID-19 provided by the risk allele was similar to, or larger than, established clinical risk factors in individuals 60 years or younger. Other common diseases have also demonstrated larger effects of genetic risk factors at a younger age (10, 11, 13). Genetic risk factors are often clinically valuable for risk stratification in younger age groups because the frequency of other established risk factors for COVID-19, such as DM, is often reduced, while the frequency of the genetic variant remains high. Moreover, this specific variant is not associated with any known COVID-19 risk factor (16) and therefore provides orthogonal information

compared with existing risk assessment tools. Although vaccination development for SARS-CoV-2 has successfully reduced COVID-19 disease burden in many countries (26, 27), SARS-CoV-2 will likely become endemic in the human population, and it is still not known how long vaccine protection will last. Therefore, this genetic variant may aid in future public health strategies, including selecting individuals for early therapy and potentially for subsequent vaccination prioritization programs.

A PRS for COVID-19 severity derived from release 6 of the COVID-19 HGI had a stronger association with COVID-19 outcomes compared with the rs10490770 risk allele alone. Neverthess, the aim of this study is to explore the clinical implications of the major genetic risk factors of COVID-19, and future studies should investigate the role of PRSs in COVID-19 severity prediction.

The biology of how the chromosome 3 genetic risk has an effect on COVID-19 severity is still unknown. This locus on chromosome 3p21 includes the putative SARS-CoV-2 coreceptors *SCL6A20* (28, 29), *LZTFL1*, and *FYCO1* (30) and the chemokine receptors *CCR9* (29), *CXCR6* (31), and *XCR1*. There are other chemokine receptors among flanking genes, *CCR1, CCR2*, and *CCR3* (32–34), whose involvement in SARS-CoV-2 infection has been suggested and could explain the biology of the striking effect of this genetic risk. Many studies (15, 29) have been trying to pinpoint a single gene or a set of causal genes, but a robust biological consensus has not been built to date.

This study has important limitations. Each cohort has its own selection bias and ascertainment bias. Several studies were enriched for severe patients, whereas UKB is a non–COVID-19 cohort, with evidence of healthy volunteer bias (35). Nevertheless, it may be less prone to

selection bias than the COVID-19 cohorts. Selection bias is inherent to most COVID-19 observational studies (36), and this influences the generalizability of the results outside the study populations. Indeed, the estimated protective effects of smoking for COVID-19 severity likely reflect the collider bias due to selection of study participants. Further, other COVID-19 epidemiological studies demonstrated similar effects (36, 37). To mitigate against these issues, we combined data from observational studies with different ascertainment strategies, including national healthcare systems, studies that were established prior to the COVID-19 pandemic so that recruitment was not dependent upon COVID-19 status, and hospital-based studies. This allowed for an increased representation of individuals with severe COVID-19 outcomes. We also provide analyses restricted to hospitalized patients, which is an ascertained, but clinically relevant, population. Although we were motivated to estimate whether homozygous individuals were at greater risk than heterozygous carriers, we could not draw any meaningful conclusions due to the low sample size (n = 135 homozygous carriers, of whom 92 were of European ancestry). While we included information from participants who were of non-European ancestry, ongoing efforts should enable larger sample sizes to better define the importance of the chromosome 3 risk locus in these ancestries. This further emphasizes the importance of developing genomics-enabled studies in individuals of non-European ancestry.

Since the beginning of the pandemic, we aimed to aggregate and harmonize individual-level clinical and genotype data from multiple cohorts from diverse countries. Due to the nature of the heterogeneity of health care systems, our data from multiple countries substantially increases the generalizability of our research findings (38). Moreover, we deposited a subset of this harmonized data to the European Genome-Phenome Archive (EGAS00001005304) for future use by all bona fide researchers to further improve our ability to understand the COVID-19 pandemic.

In summary, the major genetic COVID-19 risk locus is common and has moderate to large effects on COVID-19 outcomes, including mortality. These effects are age dependent, such that the magnitude of risk increases in younger individuals. These findings suggest potential implications of genetic information in clinical risk management.

## 3.6 Methods

### 3.6.1 Study participants.

We gathered clinical and genomic data from 13,888 COVID-19 cases (7,185 of whom were hospitalized) with genetic information available, harmonizing individual-level data from 17 studies. COVID-19 cases were defined as individuals having at least 1 confirmed SARS-CoV-2 viral nucleic acid amplification test from relevant biologic fluids or whose SARS-CoV-2 status was confirmed by ICD-10 codes, using codes U071 and/or U072. We combined data from hospital-based studies that recruited participants after COVID-19 outbreak and a population-based biobank in which recruitment was not dependent upon COVID-19 status. Data were centrally collected at the Institute for Molecular Medicine Finland and harmonized through a standardized data dictionary (https://docs.google.com/spreadsheets/d/1hwBeqckB3_qC8nnavT0kLLntOh3GrmWRJQHeO9zwG8w/edit#gid=665246845). Detailed information for data collection in each individual study is described in Supplemental Methods.

### 3.6.2 Genotyping and ancestry assignment

In order to tag the chromosome 3 locus, we selected the SNP rs10490770, which was most significantly associated with hospitalization in the COVID-19 GWAS from the COVID-19 HGI, since this is the largest GWAS metaanalysis of COVID-19 severity (ref. [16];

cases/controls = 12,888/1,295,966). Each participating study used genotyping and imputation separately following a recommended quality control pipeline (https://docs.google.com/document/d/16ethjgi4MzlQeO0KAW_yDYyUHdB9k-KbtfuGW4XYVKQg/edit). Detailed methods describing genotyping and imputation are available in Supplemental Methods. Ancestry was inferred by performing projection onto the PC analysis (PCA) space from the 1000G (20) phase 3 population using HapMap3 SNPs (39) with minor allele frequency greater than 1% (detailed methods are in Supplemental Methods; Supplemental Table 16 and Supplemental Figure 1).

### 3.6.3 Statistical analyses

To test the association between rs10490770 and all phenotypes, we applied a dominant model by grouping participants into 2 groups according to their genotype at rs10490770. C is the allele associated with COVID-19 severity; those with TC genotype or CC genotype were labeled as carriers, and those with TT genotype were labeled as noncarriers. We chose this model because it had the lowest Akaike information criterion (AIC) compared with additive and recessive models (see the Supplemental Methods and Supplemental Table 17 for details) for a logistic regression for death or severe respiratory failure outcome (defined below). All analyses were performed separately for each ancestry group. Because the sample size in non-Europeans was limited, we reported the results from individuals of European descent as the main analyses, but the results from non-European ancestry individuals are in Supplemental Figures 4–7. All analyses were based on mixed-effects models adjusted for age and sex, and the first 5 genetic PCs as fixed effects and study groups were also included as random effects to account for study variability. Five study groups, mostly reflecting the country of origin of the study, were created by combining small participating studies with few cases and controls to reduce the risk of collinearity (details are described in Supplemental Methods. We further estimated the frequency of rs10490770 risk allele carrier status from population frequencies

reported in an external database (the Genome Aggregation Database, version 3.1 [gnomAD]; ref. [19]), assuming this variant follows Hardy-Weinberg equilibrium.

### 3.6.4 Association with mortality

The HR for all-cause mortality was estimated by Cox's proportional hazard models using the coxme version 2.2-16 R package (https://cran.r-project.org/web/packages/coxme/). Individuals entered follow-up when diagnosed with COVID-19 or, if a diagnosis date was missing, when hospitalization occurred or when symptoms started. Date of death was considered an event, and data were censored at the last date of follow-up (details are described in Supplemental Methods). We additionally performed competing risk analyses to estimate the subdistribution HR for COVID-19–related mortality using the cmprsk version 2.2-10 R package, which accounts for the competing risk of non-COVID-19–related death: i.e., individuals who did not die of COVID-19 but died due to other causes (e.g., cancer). In the competing risk model, study groups were considered as fixed effects. Survival analyses were restricted to study participants with available follow-up and cause of death information (n = 9699). Cause of death was defined by doctor diagnoses, medical chart reviews, or ICD-10 codes (details are described in Supplemental Methods).

### 3.6.5 Association with COVID-19 severity and complications.

To understand the clinical implications of the chromosome 3 locus, we fit mixed-effects regression models to assess the association of rs10490770 risk allele (C) carrier status with 3 types of COVID-19–related measurements: COVID-19 severity, COVID-19 complications, and laboratory values. To do so, we defined 3 COVID-19 severity outcomes, with appropriate control definitions among SARS-CoV-2–positive individuals: (a) hospitalization; (b) ICU admission, and (c) death or severe respiratory failure. Hospitalization cases were COVID-19 cases admitted to the hospital (corresponding to WHO clinical progression scale [ref. [40]] $\geq 4$;

Supplemental Table 18), whereas controls were individuals who did not experience hospitalization (corresponding to WHO clinical progression scale [ref. [40]] 1 to 3; Supplemental Table 18). ICU cases were those COVID-19 cases admitted to the ICU, and controls were individuals who did not experience hospitalization. To assess potential selection bias, we also repeated the analyses using only individuals who were hospitalized. In these analyses, controls were defined as those who were hospitalized, but not admitted to the ICU. Death or severe respiratory failure cases were defined as individuals who died or required respiratory support (intubation, continuous positive airway pressure, bilevel positive airway pressure, or continuous external negative pressure, high-flow positive end expiratory pressure oxygen), had ICD-10 codes for acute respiratory distress syndrome (ARDS) or acute respiratory failure (J80, J9600, J9609, Z991), or OPCS codes for the use of a ventilator (E851, E852), corresponding to WHO clinical progression scale (40) ≥6 (Supplemental Table 18).

We next defined 5 COVID-19–related complications, which were diagnosed in the hospital. These included the following: (a) severe respiratory failure, which was defined as individuals who used respiratory support or had administrative codes for ARDS, respiratory failure, or ventilatory support, as described above, corresponding to WHO clinical progression scale (40) 6 to 9 (Supplemental Table 18); (b) hepatic injury, which was defined as individuals with at least 1 of the following: doctor-diagnosed hepatic complications, highest alanine aminotransferase over 3 times the upper limit of normal (ULN), or ICD-10 codes for acute hepatic failure (K720); (c) cardiovascular complications, which were defined by at least 1 of the following: doctor-diagnosed acute myocardial infarction (AMI) or stroke, highest troponin T or troponin I greater than ULN, or ICD-10 codes for AMI or stroke (I21*, I61, I62, I63, I64, I65, I66*); (d) kidney injury, defined by at least 1 of the following: doctor-diagnosed acute kidney injury (AKI), highest creatinine greater than 1.5 times ULN, or ICD-10 codes

for AKI (N17*); and (e) VTE, defined by at least 1 of the following: doctor-diagnosed pulmonary embolism (PE) or deep venous thrombosis (DVT) or ICD-10 codes for PE or DVT (I26*, I81, I82*). Controls for severe respiratory failure were defined as those requiring no oxygen therapy and who were alive, corresponding to WHO clinical progression scale (40) 1 to 4 (Supplemental Table 18), whereas controls for other complications were defined as those who did not meet the corresponding case criteria and were alive.

Finally, we considered the laboratory values of complete blood count and biochemistry tests available at hospitals (Supplemental Table 6). To test the association with the chromosome 3 locus, we used the lowest value for lymphocyte counts and otherwise the highest value recorded per individual (21–25). This is because we were interested in using these laboratory values as a proxy for COVID-19 severity. Definitions and quality control of laboratory values and specific codes are described in Supplemental Methods and Supplemental Figure 2.

#### 3.6.6 Age-dependent associations with COVID-19 severity.

We evaluated the age-dependent effects of the risk allele carrier status on the 3 COVID-19 severity phenotypes we defined above by performing 2 sets of analyses: (a) linear regressions between age at diagnosis and risk allele carrier status among severe cases, adjusting for the same covariates as the main analyses, and (b) adding a carrier status by age interaction term in the main regression models. Age was not dichotomized in these analyses. We also stratified participants by age 60 years or less or more than 60 years and repeated the same logistic regressions, and we estimated the frequency of the risk allele carriers in the 2 age groups. We used 60 years as a cut-point for age-stratified analyses because COVID-19 case fatality rates increase markedly after this age (https://www.inspq.qc.ca/covid-19/donnees/age-sexe) (41).

### 3.6.7 Associations with COVID-19 severity stratified by established clinical risk factors.

In order to compare the association of rs10490770 risk allele carrier status with other risk factors, we similarly stratified participants by BMI of 30 kg/m$^2$ or more (a definition of obesity; ref.42), smoking (ever smoker vs. never smoker), cancer, chronic kidney disease, COPD, chronic heart failure, transplantation, and DM, all of which were curated as established clinical risk factors for severe illness of COVID-19 according to the CDC website (42). All of the 8 risk factors were defined by doctor diagnoses, medical chart reviews, or ICD-10 codes (details are described in Supplemental Methods and Supplemental Table 19). We then tested the difference of the magnitude of the associations of the risk allele carrier status compared with the 8 clinical risk factors. Clinical risk factor–stratified analysis and prediction assessment (described below) were restricted to individuals with complete information for demographics, clinical risk factors, and rs10490770 genotype information (n = 7983). The majority of this subset were from UKB (n = 7461), and only 145 individuals were included from the first discovery GWAS (14).

### 3.6.8 Risk prediction compared with established clinical risk factors.

To better understand the prediction improvement by addition of the chromosome 3 genetic risk in addition to the 8 clinical risk factors, we performed multivariable regressions in individuals with complete information as described above (n = 7983). We evaluated whether the rs10490770 risk allele improved the risk prediction discrimination for severe COVID-19 outcomes by calculating the AUC and the continuous net reclassification improvement (NRI) using pROC, version 1.16.2 (https://cran.r-project.org/web/packages/pROC/index.html), and PredictABEL, version 1.2-4 R packages (https://cran.r-project.org/web/packages/PredictABEL/index.html).

### 3.6.9 Metaanalyses.

As secondary analyses, we metaanalyzed the results for non-European ancestries and 2 external cohorts for which we did not have access to individual-level data: FinnGen and Columbia University COVID-19 Biobank (CUB). This resulted in a total study population of 15,064 individuals with COVID-19. Inverse-variance weighted metaanalyses were performed under a fixed effect and random effects model using the meta version 4.16-1 R package when the appropriate phenotypes were available and case counts, control counts, and the rs10490770 risk allele carrier counts were larger than 10 in each cohort.

### 3.6.10 Sensitivity analysis.

Adjusting for participating studies may lead to reduced statistical power, given that some studies had only severe cases or had disproportional case-control ratios. To alleviate the collinearity issue, we grouped some small studies to account for study variability. This may not fully account for between-study variability. Thus, we performed 2 sets of sensitivity analyses where we included (a) only 5 genetic PCs without including the study of origin as random or fixed effects and (b) all participating studies either as fixed or random effects. Next, we performed the same analyses using UKB to provide estimates that are more representative of the general population, since this is not a COVID-19–specific cohort. We also tried binning by different cutoffs for age-stratified analyses. In order to understand whether results could have been influenced by related individuals within the samples, we selected 1 individual from a pair of relatives with PI-HAT (proportion of identity by descent calculated by PLINK; ref. [43]) greater than 0.1875 (meaning between second and third-degree relatives) and repeated the main analyses.

### 3.6.11 Statistics.

To test the association between rs10490770 and all phenotypes, we applied a dominant model by grouping participants into 2 groups according to their genotypes at rs10490770. C is the allele associated with COVID-19 severity; those with TC genotype or CC genotype were labeled as carriers, and those with TT genotype were labeled as noncarriers. All analyses were based on mixed-effects models adjusted for age, sex, and the first 5 genetic PCs as fixed effects. Study groups were also included as random effects to account for study variability. Five study groups, mostly reflecting the country of origin of the study, were created by combining small participating studies with few cases and controls to reduce the risk of collinearity. We did not apply a multiple-testing correction, and a P value of less than 0.05 was considered significant, since all the outcomes tested were related to COVID-19 severity and not independent of each other.

### 3.6.12 Data and materials availability.

All code for data management and analysis is archived online at https://github.com/tomoconaka/ COVID19-chr3 (commit 183ddb7) for review and reuse. The harmonized individual-level data of some participating cohorts from Belgium (BeLCovid_2), Brazil (BRACOVID), Italy (COVID19-Host(a) ge_4, GEN-COVID), Spain (COVID19-Host(a)ge_1,2,3, INMUNGEN- CoV2, Determining the Molecular Pathways and Genetic Predisposition of the Acute Inflammatory Process Caused by SARS-CoV-2 [SPGRX]), and Sweden (SweCovid) were deposited at the European Genome-Phenome Archive (EGA EGAS00001005304). Regarding the SweCovid study, an institutional data transfer agreement can be established and data may be shared if the aims of data use are covered by ethical approval and patient consent. Regarding the data from genetic modifiers for COVID-19–related illness (BelCovid_1), individual-level data were acquired and shared with FIMM during the early stages of the pandemic Upon contact with Isabelle Migeotte (Isabelle.

Migeotte@erasme.ulb.ac.be), an institutional data transfer agreement can be established and data can be shared if the aims of data use are covered by ethical approval and patient consent. The procedure will involve an update to the ethical approval as well as review by legal departments at both institutions, and the process will typically take 2 to 4 months from initial contact.

Regarding the BoSCO study, individual-level genotype and clinical data for the purpose of this study were shared with FIMM under a legal, bilateral agreement and were specific to this particular project. Current participant consents and privacy regulations prohibit deposition of individual level data to public repositories. Upon contact with Kerstin Ludwig (kerstin.ludwig@uni-bonn.de), an institutional data transfer agreement can be established and data shared if the aims of data use are covered by ethical approval and patient consent. The procedure will involve review by legal departments at both institutions, and the process will typically take about 2 months from initial contact.

The BQC19 is an open science biobank. Instructions on how to access data for individuals from the BQC19 at the Jewish General Hospital site are available here: https://www.mcgill.ca/genepi/ mcg-covid-19-biobank. Instructions on how to access data from other sites of the BQC19 are available here: https://www.bqc19.ca/en/ access-data-samples.

For the COVID-19 Kohortenstudie am Klinikum München Rechts der Isar (COMRI) cohort, data protection legislation does not allow for deposition of individual level data in public repositories. Upon direct contact with Christoph Spinner (christoph.spinner@tum.de), an institutional data transfer agreement can be established and data will be shared if the aims of data use are covered by ethical approval and patient consent. The procedure will involve an

update to the ethical approval as well as review by legal departments at both institutions, and the process will typically take 2 to 3 months from initial contact.

Regarding the Fondazione IRCCS Milan data (FOGS study), institutional data privacy regulations prohibit deposition of individual level data to public repositories without specific consent. Participant written consent also does not cover public sharing of data for use for unknown purposes. Upon contact with Luca Valenti (luca.valenti@ unimi.it), an institutional data transfer agreement can be established and data shared if the aims of data use are covered by ethical approval and patient consent. The procedure will involve the request for an amendment to the ethical approval as well as review by legal departments at both institutions, and the process will typically take 1 to 2 months from initial contact.

Regarding Norwegian data (the Norwegian SARS-CoV-2 study), institutional data privacy regulations prohibit deposition of individual level data to public repositories. Participant written consent also does not cover public sharing of data for use for unknown purposes. Upon contact with Tom H. Karlsen (t.h.karlsen@medisin.uio.no) or Johannes R. Hov (j.e.r.hov@medisin.uio.no), an institutional data transfer agreement can be established and data shared if the aims ofdata use are covered by ethical approval and patient consent. The procedure will involve an update to the ethical approval as well as review by legal departments at both institutions, and the process will typically take 1 to 2 months from initial contact.

The genetic and phenotype data sets from UKB are available via the UKB data access process (see http://www.ukbiobank.ac.uk/ register-apply/).

### 3.7 Note added in proof.

Using chromosome conformation capture and gene-expression analysis, a recent study identified the gain-of-function SNP for *LZTFL1*, rs17713054G>A, as a probable causative variant conferring increased risk of respiratory failure with COVID-19 (44)

## 3.8 Figures

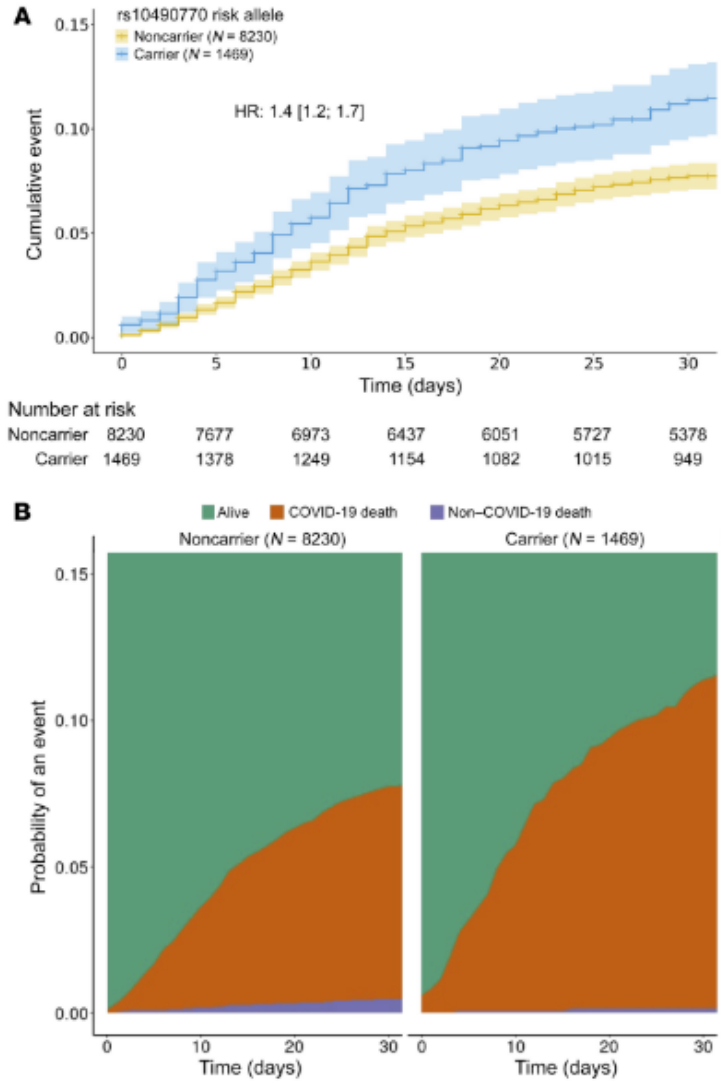


**Figure 1. Associations with mortality.**

The results described here were restricted to 9699 COVID-19 patients of European ancestry with available follow-up and cause of death information. (A) Survival analysis using Cox's proportional hazard model. Kaplan-Meier curves stratified by rs10490770 risk allele carrier status. (carriers: n = 1469 vs. noncarriers: n = 8,230). HRs were calculated by adjusting for age, sex, and genetic PCs 1 to 5 as fixed effects and a dummy variable representing the participating studies as random effects. (B) Cumulative incidence curves for COVID-19–related death and COVID-19–unrelated death among the same individuals as described in A.

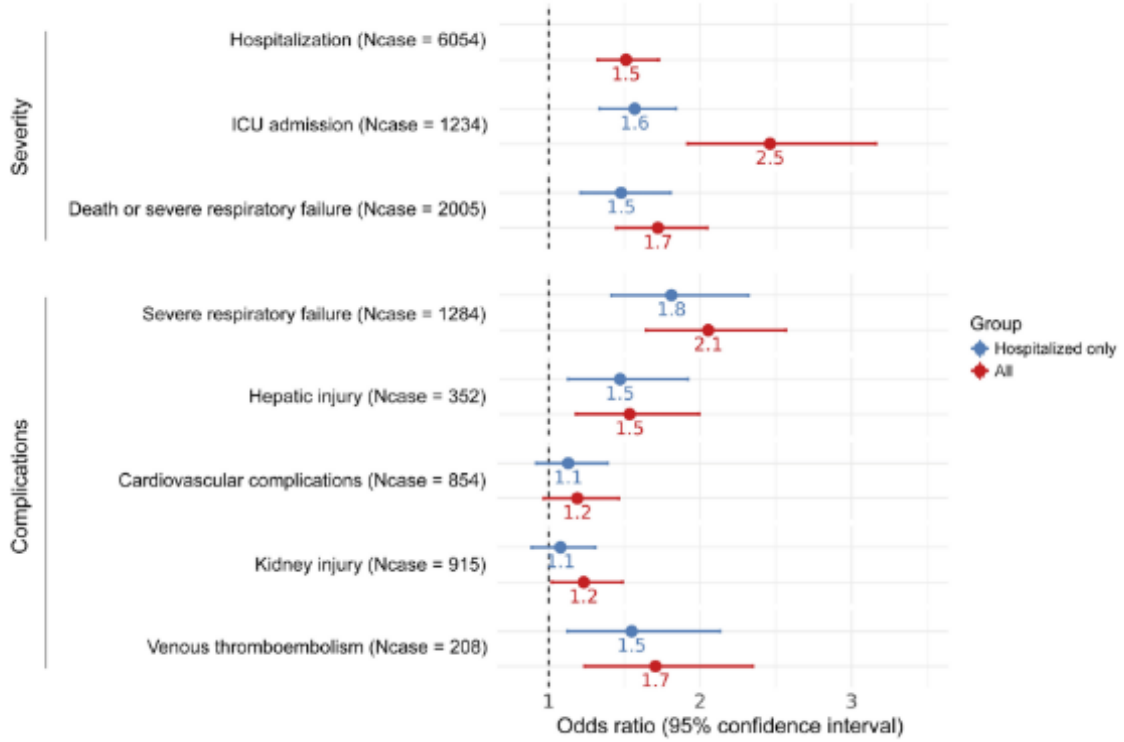


**Figure 2. Associations between rs10490770 risk allele carrier status and COVID-19 severity and complications.**

The results described here were restricted to COVID-19 patients of European ancestry. Logistic regressions were fit to assess the associations of rs10490770 risk allele carrier status with COVID-19 severity and complications, adjusting for age, sex, and genetic PCs 1 to 5 as fixed effects, and a dummy variable representing the participating studies as random effects. Red: All participants (n = 12,091); blue: hospitalized participants only (n = 6054). The case counts demonstrated as Ncase are the case counts in the analyses of all participants. The full list of case and control counts in the analyses of all participants and those hospitalized only are described in Supplemental Table 5.

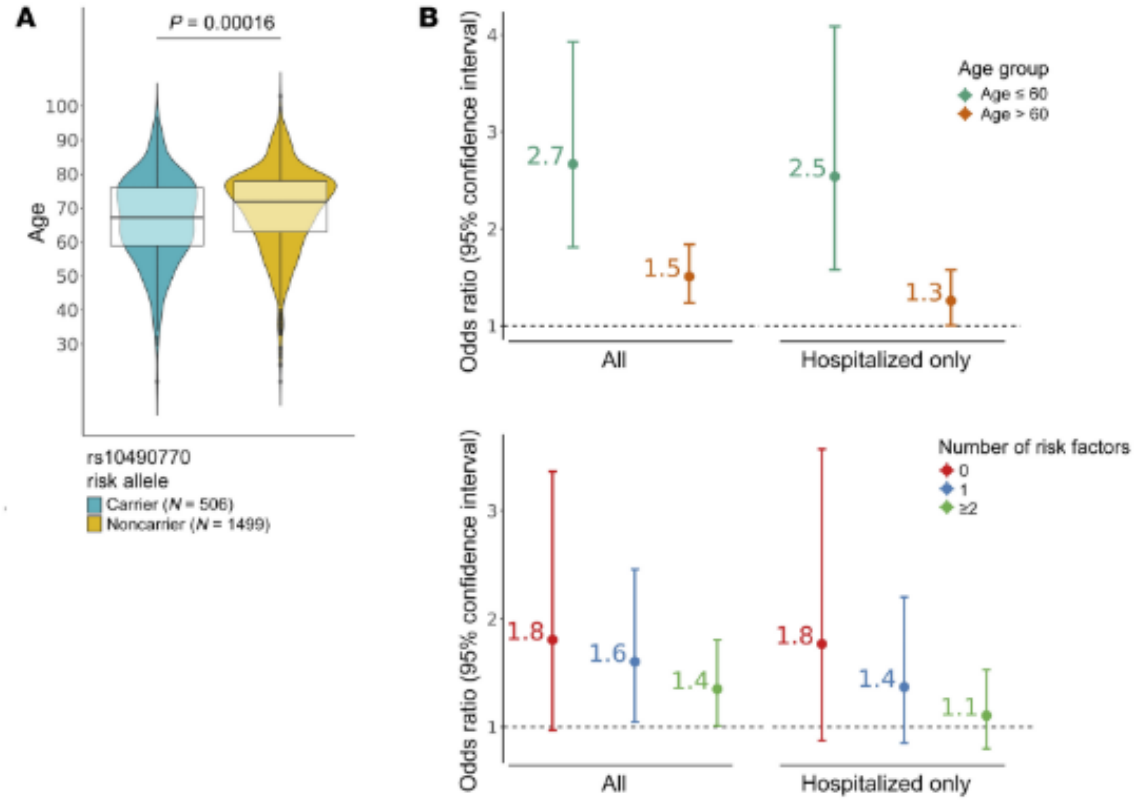


**Figure 3. Influence of age and clinical risk factors for the effect of rs10490770 risk allele carrier status on death or severe respiratory failure.**

(A) Age distribution in COVID-19 patients of European ancestry who died or experienced severe respiratory failure (n = 2005). Median (IQR) age was 67.2 (range, 59–76) years in carriers (n = 506) and 72 (range, 63–78) years in noncarriers (n = 1499). (B) ORs of rs10490770 risk allele carrier status for death or severe respiratory failure. Regressions were performed within subgroups stratified by age (age ≤60 years and age >60 years) (cases/controls = 2005/7047) or by the number of established risk factors (0, 1, or ≥2); BMI ≥30, smoking, cancer, chronic kidney disease, COPD, chronic heart failure, transplantation, and DM (cases/controls = 898/6454). All analyses were adjusted for age, sex, genetic PCs 1 to 5 as fixed effects, and a dummy variable representing the participating studies as random effects.

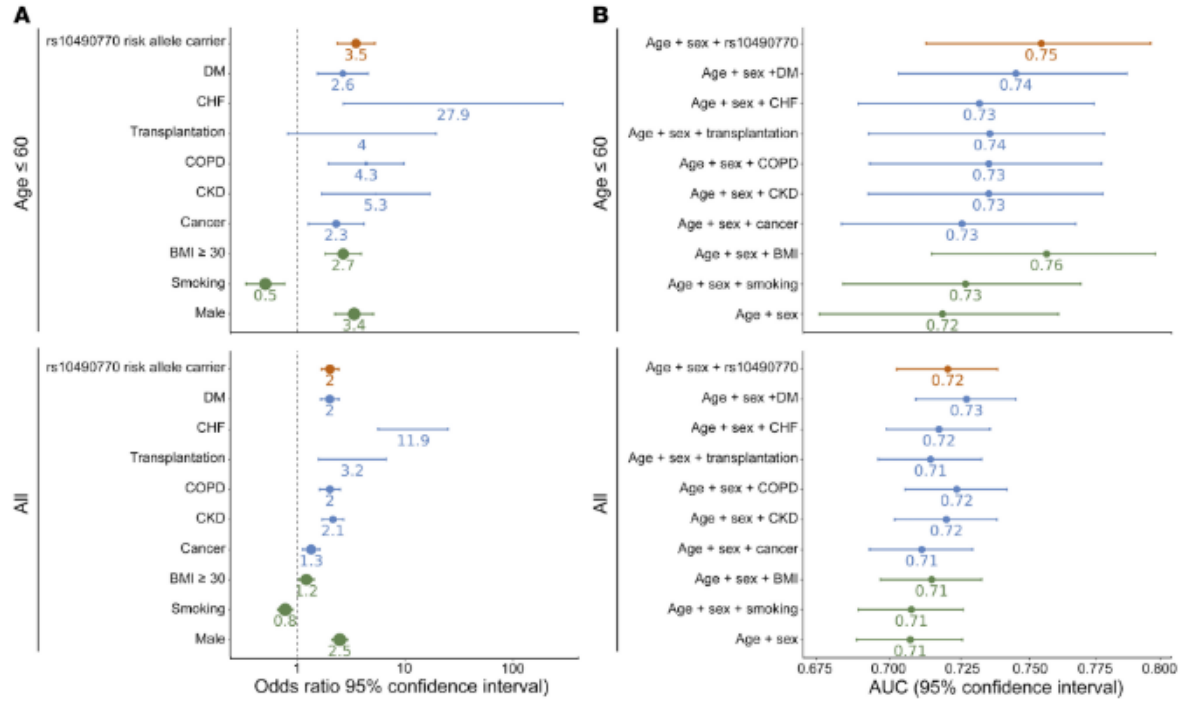


**Figure 4. Multivariable regression models and risk prediction estimates for death or severe respiratory failure.**

Multivariable regression analyses for death or severe respiratory failure were restricted to European-ancestry individuals with complete information of demographic variables (green), comor- bidities (blue), and rs10490770 risk allele status (red). n = 7352 for all and n = 2499 for age ≤60. CKD, chronic kidney disease; CHF, chronic heart failure. Error bars indicate 95% CIs. (A) Forest plots comparing ORs from multivariable regression models. The size of each dot represents the frequency of the risk factors. (B) Comparison of AUCs of predictions for COVID-19 outcomes. rs10490770 risk allele and nongenetic clinical risk factors were included separately in addition to age and sex in multivariable regression models.

### 3.9 Tables

**Table 1. Participant characteristics.**

| | Hospitalized | Total |
|---|---|---|
| | (*n* = 7185) | (*n* = 13,888) |
| Female | 2866 (39.9%) | 6549 (47.2%) |
| Age (years)[A] | 64.8 (14.7) | 63.7 (12.8) |
| Ancestry | | |
| European | 6054 (84.3%) | 12,091 (87.1%) |
| South Asian | 113 (1.6%) | 389 (2.8%) |
| African | 234 (3.3%) | 421 (3.0%) |
| Others | 187 (2.6%) | 276 (2.0%) |
| East Asian | 64 (0.9%) | 109 (0.8%) |
| Admixed American | 533 (7.4%) | 602 (4.3%) |
| ICU admission | 1695 (24.3%) | 1695 (12.5%) |
| Death status | | |
| Survived | 4887 (79.3%) | 11,369 (90.0%) |
| Deceased | 1264 (20.5%) | 1264 (10.0%) |
| Respiratory failure | | |
| Severe respiratory failure | 1704 (30.2%) | 1704 (14.6%) |
| Oxygen supplementation | 2051 (36.4%) | 2051 (17.6%) |
| Hepatic injury | 532 (10.8%) | 536 (4.7%) |
| Cardiovascular complications | 1017 (19.6%) | 1040 (9.3%) |
| Kidney injury | 1172 (21.8%) | 1182 (10.0%) |
| VTE | 288 (6.9%) | 289 (2.7%) |

[A]Mean (SD); percentage was calculated among those with complete information. The missing rates for each study are listed in Supplemental Table 1. Others in ancestry included remaining individuals who were not assigned as either of European, South Asian, African, East Asian, or admixed American descent.

**Table 2. Age and risk allele carrier status by COVID-19 severity outcomes.**

| | Death or severe respiratory failure | COVID positive but no oxygen supplementation | |
|---|---|---|---|
| | | Hospitalized | All |
| All | | | |
| Carrier | 25.2% [23.4; 27.2] (506) | 16.2% [14.5; 18.1] (261) | 13.8% [13; 14.6] (974) |
| Noncarrier | 74.8% [72.8; 76.6] (1499) | 83.8% [81.9; 85.5] (1346) | 86.2% [85.4; 87] (6073) |
| Total | 100% (2,005) | 100% (1607) | 100% (7047) |
| Age ≤60 years old | | | |
| Carrier | 32.3% [28.3; 36.7] (151) | 14.6% [11.3; 18.7] (52) | 13.9% [12.6; 15.2] (366) |
| Noncarrier | 67.7% [63.3; 71.7] (316) | 85.4% [81.3; 88.7] (304) | 86.1% [84.8; 87.4] (2274) |
| Total | 100% (467) | 100% (356) | 100% (2640) |
| Age >60 years old | | | |
| Carrier | 23.1% [21; 25.3] (355) | 16.7% [14.7; 18.9] (209) | 13.8% [12.8; 14.8] (608) |
| Noncarrier | 76.9% [74.7; 79] (1183) | 83.3% [81.1; 85.3] (1042) | 86.2% [85.2; 87.2] (3799) |
| Total | 100% (1538) | 100% (1251) | 100% (4407) |

Frequency of rs10490770 risk variant carriers in individuals of European descent stratified by age and COVID-19 severe outcomes. Square brackets indicate 95% CI; parentheses show sample size.

**Table 3. Risk prediction performance for death or severe respiratory failure.**

| Age range | Model | AUC[A] | AUC *P* value[B] | NRI[A] | NRI *P* value[B] |
|---|---|---|---|---|---|
| All<br>Cases = 898<br>Controls = 6454 | Baseline | 0.76 [0.75; 0.78] | – | – | – |
| | Baseline and rs10490770 | 0.77 [0.76; 0.79] | $1.4 \times 10^{-4}$ | 0.19 [0.13; 0.25] | $4.4 \times 10^{-11}$ |
| Age ≤60<br>Cases = 151<br>Controls = 2,348 | Baseline | 0.82 [0.79; 0.86] | – | – | – |
| | Baseline and rs10490770 | 0.84 [0.81; 0.88] | $2.1 \times 10^{-2}$ | 0.41 [0.26; 0.56] | $7.7 \times 10^{-8}$ |

Only individuals with complete information regarding clinical risk factors and genotype were included. Baseline model includes age, sex, BMI, smoking status (ever smoker versus never smoker), cancer, chronic kidney disease, COPD, chronic heart failure, transplantation, and DM. [A]Square brackets show 95% CI. [B]P values were calculated by comparing baseline model and baseline and rs10490770 model.

### 3.10 List of references


1. McKee M, Stuckler D. If the world fails to protect the economy, COVID-19 will damage health not just now but also in the future. *Nat Med*. 2020;26(5):640–642.

2. Buitrago-Garcia D, *et al*. Occurrence and transmission potential of asymptomatic and presymptomatic SARSCoV-2 infections: A living systematic review and meta-analysis. *PLoS Med*. 2020;17(9):e1003346.

3. Dooling K, *et al*. The Advisory Committee on immunization practices' updated interim recom- mendation for allocation of COVID-19 vaccine - United States, December 2020. *MMWR Morb Mortal Wkly Rep*. 2021;69(5152):1657–1660.

4. Bubar KM, *et al*. Model-informed COVID-19 vac- cine prioritization strategies by age and serostatus. *Science*. 2021;371(6532):916–921.

5. Novelli G, *et al*. COVID-19 one year into the pandemic: from genetics and genomics to therapy, vaccination, and policy. *Hum Genomics*. 2021;15(1):1–13.

6. O'Driscoll M, *et al*. Age-specific mortality and immunity patterns of SARS-CoV-2. *Nature*. 2020;590(7844):140–145.

7. Williamson EJ, *et al*. Factors associated with COVID-19-related death using OpenSAFELY. *Nature*. 2020;584(7821):430–436.

8. Chaudhry R, *et al*. A country level analysis measuring the impact of government actions, country preparedness and socioeconomic factors on COVID-19 mortality and related health outcomes. *EClinicalMedicine*. 2020;25:100464.

9. Baric RS. Emergence of a highly fit SARS-CoV-2 variant. *N Engl J Med*. 2020;383(27):2684–2686.

10. Van Der Kolk DM, *et al*. Penetrance of breast can- cer, ovarian cancer and contralateral breast cancer in *BRCA1* and *BRCA2* families: High cancer incidence at older age. *Breast Cancer Res Treat*. 2010;124(3):643–651.

11. Nordestgaard BG, *et al*. Familial hypercholesterolaemia is underdiagnosed and undertreated in the general population: guidance for clinicians to prevent coronary heart disease: Consensus Statement of the European Atherosclerosis Society. *Eur Heart J*. 2013;34(45):3478–3490.

12. Feng YCA, *et al*. Findings and insights from the genetic investigation of age of first reported occurrence for complex disorders in the UK Biobank and FinnGen [preprint]. https://doi.org/10. 1101/2020.11.20.20234302. Posted on *medRxiv* November 25, 2020.

13. Olarte L, *et al*. Apolipoprotein E epsilon4 and age at onset of sporadic and familial Alzheimer disease in Caribbean Hispanics. *Arch Neurol*. 2006;63(11):1586–1590.

14. The Severe Covid-19 GWAS Group. Genome wide association study of severe Covid-19 with respiratory failure. *N Engl J Med*. 2020;383:1522–1534.

15. Pairo-Castineira E, *et al*. Genetic mechanisms of critical illness in COVID-19. *Nature*. 2020;591(7848):92–98.

16. COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19 [published online July 8, 2021]. *Nature*. https://doi. org/10.1038/s41586-021-03767-x.

17. Kosmicki JA, *et al*. Pan-ancestry exome-wide association analyses of COVID-19 outcomes in 586,157 individuals. *Am J Hum Genet*. 2021;108(7):1350–1355.

18. Zeberg H, Pääbo S. The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. *Nature*. 2020;587(7835):610–612.

19. Karczewski KJ, *et al*. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581(7809):434–443.

20. Auton A, *et al*. A global reference for human genetic variation. *Nature*. 2015;526(7571):68–74.

21. Del Valle DM, *et al*. An inflammatory cytokine signature predicts COVID-19 severity and survival. *Nat Med*. 2020;(3):1–8.

22. Merrill JT, *et al*. Emerging evidence of a COVID- 19 thrombotic syndrome has treatment implica- tions. *Nat Rev Rheumatol*. 2020;16(10):581–589.

23. Higuera-de la Tijera F, *et al*. Impact of liver enzymes on SARS-CoV-2 infection and the severity of clinical course of COVID-19. *Liver Res*. 2021;5(1):21–27.

24. Vafadar Moradi E, *et al*. Increased age, neutrophil-to-lymphocyte ratio (NLR) and white blood cells count are associated with higher COVID-19 mortality. *Am J Emerg Med*. 2021;40:11–14.

25. Yan L, *et al*. An interpretable mortality prediction model for COVID-19 patients. *Nat Mach Intell.* 2020;2(5):283–288.

26. Polack FP, *et al*. Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. *N Engl J Med*. 2020;383(27):2603–2615.

27. Poustchi H, *et al*. SARS-CoV-2 antibody seroprevalence in the general population and high-risk occupational groups across 18 cities in Iran: a population-based cross-sectional study. *Lancet Infect Dis*. 2021;21(4):473–481.

28. Vuille-Dit-Bille RN, *et al*. Human intestine luminal *ACE2* and amino acid transporter expression increased by ACE-inhibitors. *Amino Acids*. 2015;47(4):693–705.

29. Yao Y, *et al*. Genome and epigenome editing identify *CCR9* and *SLC6A20* as target genes at the 3p21.31 locus associated with severe COVID- 19. *Signal Transduct Target Ther*. 2021;6(1):85.

30. Smieszek SP, Polymeropoulos MH. Role of FYVE and coiled-coil domain autophagy adaptor 1 in severity of COVID-19 1 infection 2 [preprint]. https://doi.org/10.1101/2021.01.22.21250070. Posted on *medRxiv* January 27, 2021.

31. Payne DJ, *et al*. The CXCR6/CXCL16 axis links inflamm-aging to disease severity in COVID-19 patients [preprint]. https://doi. org/10.1101/2021.01.25.428125. Posted on *bioRxiv* January 25, 2021.

32. Khalil BA, *et al*. Chemokines and chemokine receptors during COVID-19 infection. *Comput Struct Biotechnol J*. 2021;19:976–988.

33. Chua RL, *et al*. COVID-19 severity correlates with airway epithelium-immune cell interactions identified by single-cell analysis. *Nat Biotechnol*. 2020;38(8):970–979.

34. Liao M, *et al*. Single-cell landscape of bronchoal- veolar immune cells in patients with COVID-19. *Nat Med*. 2020;26(6):842–844.

35. Fry A, *et al*. Comparison of sociodemographic and health-related characteristics of UK biobank participants with those of the general population. *Am J Epidemiol*. 2017;186(9):1026–1034.

36. Griffith GJ, *et al*. Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nat Commun*. 2020;11(1):1–12.

37. Miyara M, *et al*. Low incidence of daily active tobacco smoking in patients with symptomatic COVID-19 [preprint]. https://doi.org/10.32388/ WPP19W.3. Posted on Qeios April 21, 2020.

38. Di Maria E, *et al*. Genetic variants of the human host influencing the coronavirus-associated phenotypes (SARS, MERS and COVID-19): rapid systematic review and field synopsis. *Hum Genomics*. 2020;14(1):30.

39. Altshuler DM, *et al*. Integrating common and rare genetic variation in diverse human populations. *Nature*. 2010;467(7311):52–58.

40. Marshall JC, *et al*. A minimal common outcome measure set for COVID-19 clinical research. *Lancet Infect Dis*. 2020;20(8):e192–e197.

41. Signorelli C, Odone A. Age-specific COVID-19 case-fatality rate: no evidence of changes over time. *Int J Public Health*. 2020;65(8):1435–1436.

42. Centers for Disease Control and Prevention. People With Certain Medical Conditions. https:// www.cdc.gov/coronavirus/2019-ncov/need- extra-precautions/people-with-medical- conditions.html. Updated August 20, 2021. Accessed January 29, 2021.

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

44. Downes DJ, *et al*. Identification of *LZTFL1* as a candidate effector gene at a COVID-19 risk locus. *Nat Genet*. 2021;53(11):1606–1615.

### 3.11 Supplemental data

Supplementary Methods, Tables and Figures can be downloaded from the open access publication Nakanishi *et al*. in *J Clin Invest* available here:

https://www.jci.org/articles/view/152386#sd

**Connecting Text: Bridge Between Chapter 3 and Chapter 4**

In the previous Chapters (2 and 3), we assessed the magnitude of the effects of genetic risks on disease susceptibility and severity, in the context of AATD and COVID-19, respectively. For both diseases, we provided evidence that genetic information could reliably predict disease onset and/or severity, as a potential translational route to use genetic profiling of individual patterns of disease predisposition. We showcased the two examples of the clinical implications of genetics to develop more personalized approaches to disease management.

In the next two Chapters (4 and 5), we sought another translational path, which is the identification of therapeutic targets within causal pathways through MR. In Chapter 4, we performed a MR study to identify novel potentially disease-influencing proteins for IPF. IPF is a progressive, fatal fibrotic interstitial lung disease that affects adults, leading to respiratory failure with a median survival time from diagnosis of 3–5 years. Despite two antifibrotic therapies have been approved for the treatment of IPF: nintedanib and pirfenidone, which slow the decline in lung function and reduce the risk of acute respiratory deterioration, many individuals with IPF remain untreated. Although several serum biomarkers for IPF have been identified, these biomarkers still lack strong evidence of disease causality and are more useful at defining prognosis once IPF has occurred. We therefore applied MR, a causal inference technique, to identify potentially causal plasma proteins which influence the IPF susceptibility and could serve as drug targets in the future.

**Chapter 4: Genetically increased circulating FUT3 level leads to reduced risk of Idiopathic Pulmonary Fibrosis: a Mendelian Randomisation Study.**

## 4.1 Title page

**Genetically increased circulating FUT3 level leads to reduced risk of Idiopathic Pulmonary Fibrosis:a Mendelian Randomisation Study**

Tomoko Nakanishi[1–4], Agustin Cerani[2,5], Vincenzo Forgetta[2], Sirui Zhou[2,5], Richard J. Allen[6], Olivia C. Leavy[6], Masaru Koido[7], Deborah Assayag[8,9], R. Gisli Jenkins[10,11], Louise V. Wain[6,12], Ivana V. Yang[13,14], G. Mark Lathrop[15], Paul J. Wolters[16], David A. Schwartz[14,17], J. Brent Richards[1,2,5,18,19]

**Affiliations:** [1]Dept of Human Genetics, McGill University, Montréal, QC, Canada. [2]Centre for Clinical Epidemiology, Dept of Medicine, Lady Davis Institute, Jewish General Hospital, McGill University, Montréal, QC, Canada. [3]Kyoto–McGill International Collaborative School in Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan. [4]Japan Society for the Promotion of Science, Tokyo, Japan. [5]Dept of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, QC, Canada. [6]Dept of Health Sciences, University of Leicester, Leicester, UK. [7]Dept of Cancer Biology, Institute of Medical Science, The University of Tokyo, Tokyo, Japan. [8]Dept of Medicine, Faculty of Medicine, McGill University, Montréal, QC, Canada. [9]Translational Research in Respiratory Diseases, Research Institute of the McGill University Health Centre (RI-MUHC), Montréal, QC, Canada. 10National Heart and Lung Institute, Imperial College London, London, UK. [11]Dept of Interstitial Lung Disease, Royal Brompton and Harefield Hospitals, Guy's and St Thomas' NHS Foundation Trust, London, UK. [12]National Institute for Health Research, Leicester Respiratory Biomedical Research Centre, Glenfield Hospital, Leicester, UK. [13]Center for Genes, Environment and Health and Dept of Medicine, National Jewish Health, Denver, CO, USA. [14]Dept of Medicine, School of Medicine, University of Colorado Denver, Aurora, CO,

USA. [15]McGill Genome Centre and Dept of Human Genetics, McGill University, Montréal, QC, Canada. [16]Dept of Medicine, School of Medicine, University of California, San Francisco, CA, USA. [17]Dept of Immunology, School of Medicine, University of Colorado Denver, Aurora, CO, USA. [18]Division of Endocrinology, Dept of Medicine, Jewish General Hospital, McGill University, Montréal, QC, Canada. [19]Dept of Twin Research, King's College London, London, UK.

**Published in:** *Eur Respir J* 2022 59: 2003979; doi: 10.1183/13993003.03979-2020

## 4.2 Abstract

**Background** Idiopathic pulmonary fibrosis (IPF) is a progressive, fatal fibrotic interstitial lung disease. Few circulating biomarkers have been identified to have causal effects on IPF.

**Methods** To identify candidate IPF-influencing circulating proteins, we undertook an efficient screen of circulating proteins by applying a two-sample Mendelian randomisation (MR) approach with existing publicly available data. For instruments, we used genetic determinants of circulating proteins which reside *cis* to the encoded gene (*cis*-single nucleotide polymorphisms (SNPs)), identified by two genome-wide association studies (GWASs) in European individuals (3301 and 3200 subjects). We then applied MR methods to test if the levels of these circulating proteins influenced IPF susceptibility in the largest IPF GWAS (2668 cases and 8591 controls). We validated the MR results using colocalisation analyses to ensure that both the circulating proteins and IPF shared a common genetic signal.

**Results** MR analyses of 834 proteins found that a 1 sd increase in circulating galactoside 3(4)-l-fucosyltransferase (FUT3) and α-(1,3)-fucosyltransferase 5 (FUT5) was associated with a reduced risk of IPF (OR 0.81, 95% CI 0.74–0.88; $p=6.3\times10^{-7}$ and OR 0.76, 95% CI 0.68–0.86; $p=1.1\times10^{-5}$, respectively). Sensitivity analyses including multiple *cis*-SNPs provided similar estimates both for FUT3 (inverse variance weighted (IVW) OR 0.84, 95% CI 0.78–0.91; $p=9.8\times10^{-6}$ and MR-Egger OR 0.69, 95% CI 0.50–0.97; p=0.03) and FUT5 (IVW OR 0.84, 95% CI 0.77–0.92; $p=1.4\times10^{-4}$ and MR-Egger OR 0.59, 95% CI 0.38–0.90; p=0.01). FUT3 and FUT5 signals colocalised with IPF signals, with posterior probabilities of a shared genetic signal of 99.9% and 97.7%, respectively. Further transcriptomic investigations supported the protective effects of FUT3 for IPF.

**Conclusions** An efficient MR scan of 834 circulating proteins provided evidence that genetically increased circulating FUT3 level is associated with reduced risk of IPF.

## 4.3 Introduction

Idiopathic pulmonary fibrosis (IPF) is a progressive, fatal fibrotic interstitial lung disease that affects adults, leading to decreased lung compliance, disrupted gas exchange and resultant respiratory failure [1]. The median survival time from diagnosis is 3–5 years, which is worse than the prognosis of most types of cancers [2]. Early detection or prevention of IPF is important as the currently available therapies are anti-fibrotic agents that have been shown to slow disease progression [3, 4]. At present, the only way to detect early disease is through high-resolution computed tomography scanning, which reveals interstitial lung abnormalities in up to 10% of the population aged >60 years, in whom only a small minority progress to develop IPF [5]. Therefore, a serum biomarker that can predict or refine disease risk through a causal relationship is urgently required.

Although several serum biomarkers for IPF have been identified [6–9], these biomarkers still lack strong evidence of disease causality and are more useful at defining prognosis once IPF has occurred. Causal inference in IPF through traditional observational studies is challenging due to potential confounding and reverse causation that can bias estimates of the effects of biomarkers on IPF. For example, smoking, a known risk factor for IPF, is confounded by its association with many other lifestyle choices. Similarly, IPF itself may influence the level of the biomarker, a phenomenon known as reverse causation. This last source of bias is particularly difficult to rule out since the timing of IPF onset is most often unknown.

Despite these challenges, identifying IPF-influencing circulating proteins is helpful as such markers could serve as both drug targets to decrease susceptibility and noninvasive biomarkers of disease risk. One way to estimate the causality of circulating biomarkers is using Mendelian randomisation (MR), which uses germline genetic variants as instrumental variables to assess the role of risk factors in disease susceptibility. Since genetic variants are

randomly assigned at conception, this process of randomisation largely breaks the association with most confounding factors. Furthermore, since germline genetic variants are always assigned prior to disease onset, reverse causation can be avoided. A further advantage of MR studies is that they can provide an assessment of a lifetime of risk factor exposure assuming the effect of the genetic variant on the risk factor is stable throughout an individual's life [10].

The goal of this study was therefore to identify circulating proteins which influence the risk for IPF by applying a MR design that efficiently screened hundreds of proteins. Bayesian colocalisation analyses were undertaken to ensure that candidate circulating proteins and IPF shared a common aetiological genetic signal and that the MR results were not biased by linkage disequilibrium (LD). Candidate IPF-influencing proteins identified through MR and colocalisation analyses were further evaluated via literature and genetic phenotype database searches and transcriptomic investigations. The results from these experiments could provide a better understanding of the aetiology of IPF and could potentially identify targets for future therapies.

## 4.4 Material and methods

### 4.4.1 Study design and data sources

We applied a two-sample MR design to identify circulating proteins associated with risk of IPF. For this, summary data were obtained from the largest IPF genome-wide association study (GWAS) to date in individuals of European ancestry [11] and from the two protein quantitative trait loci (pQTL) GWASs by Sun *et al*. [12] and Emilsson *et al*. [13]. Detailed methods of protein assays are described in each study [12, 13]. See figure 1 for a schema of our study design.

### 4.4.2 Ethical approval

No separate ethical approval was required due to the use of publicly available data.

### 4.4.3 Mendelian randomization

MR relies upon three major assumptions [14]. First, the genetic variants must reliably associate with the exposure. With the advent of large-scale modern GWASs, genetic variants associating with exposure can be identified in large datasets [15]. Second, the genetic variants must not be associated with confounders of the exposure–outcome relationship. A potential violation of this assumption can occur due to confounding by LD and/or population ancestry [16]. Lastly, genetic variants must not affect the outcome, except through the exposure of interest (referred to as a lack of horizontal pleiotropy) [17].

Large-scale GWASs for circulating proteins [12, 13] have often found that the genetic determinants of circulating proteins reside *cis* (in close proximity) to the encoding genes. The use of *cis*-acting single nucleotide polymorphisms (SNPs) for MR reduces potential horizontal pleiotropy and increases the validity of MR assumptions, because a *cis*-SNP strongly associated with the protein is likely to directly influence the gene's transcription and consequently the circulating protein level. We selected independent ($r^2 \leq 0.001$) *cis*-pQTL SNPs that are significantly associated with circulating proteins ($p < 5 \times 10^{-8}$) from two pQTL GWASs [12, 13]. More details are provided in the supplementary material.

### 4.4.4 Statistical analysis

We performed MR using the TwoSampleMR R package [18]. For proteins with a single *cis*-SNP, the Wald estimator ($\beta_{IPF}/\beta_{protein}$) was used to estimate the effect of the protein on IPF risk. Where multiple SNPs were available, our primary analyses used an inverse variance weighted (IVW) estimator [19]. Benjamini–Hochberg correction was applied to adjust for the

multiple proteins tested, which is likely to be conservative because some protein levels are partially correlated with each other (false discovery rate 0.05 with 507 multiple testing for Sun *et al*. [12] and 733 multiple testing for Emilsson *et al*. [13]).

### 4.4.5 Colocalization analysis

Candidate IPF-influencing proteins supported by MR were evaluated via colocalisation analyses using the coloc R package [20] and eCAVIAR [21] for the proteins in Sun *et al*. [12], which provided genome-wide summary statistics for each protein. Colocalisation analysis is a way to estimate the posterior probability of whether the same genetic variants are responsible for the two GWAS signals (in this case, protein level and IPF) or they are distinct causal variants that are just in LD with each other. Detailed methods are described in the supplementary material. LocusZoom plots were created to visualise these colocalisations [22].

### 4.4.6 Sensitivity analyses

Sensitivity analyses were performed for proteins with support from MR and colocalisation analyses. Multiple *cis*-SNPs in weak LD ($r^2$<0.6) with the leading *cis*-SNPs for candidate proteins were included in IVW and MR-Egger analyses that considered correlated variants using the MendelianRandomisation R package [23, 24], because consistency of estimates could strengthen the hypothesised effects. MR-Egger allows for a y-intercept term using a random effects model. An intercept different from zero indicates directional horizontal pleiotropy, suggestive of a violation of the third MR assumption. Detailed methods are described in the supplementary material. Bidirectional MR was also conducted to test whether IPF had an effect on candidate protein levels.

To further test for the presence of horizontal pleiotropy, potential pleiotropic effects of each protein-associated SNP were searched using PhenoScanner [25, 26], a database with over 65 billion associations and over 150 million unique genetic variants.

### 4.4.7 Transcriptomic data in lung tissue

We further investigated FUT3 and FUT5 using microarray-based transcriptomic data in whole lungs: GSE32537 [27]. Logistic regression was fitted to assess the associations between IPF and standardised log-transformed expressions, adjusted for age, sex and smoking status (ever versus never). We additionally explored the expression profiles using two single-cell RNA sequencing (scRNA-seq) datasets: GSE135893 [28] and GSE136831 [29]. The unique molecular identifier counts of FUT3 were compared between IPF and control subjects, stratified by each cell type annotation according to the original publications. Detailed methods are described in the supplementary material.

## 4.5 Results

### 4.5.1 Cohort characteristics

The GWAS of circulating protein levels from the INTERVAL study by Sun *et al*. [12] consisted of 3301 participants of European descent in England (mean age 43.7 years) (table 1). The circulating protein GWAS from the AGES Reykjavik study by Emilsson *et al*. [13] recruited 3200 Icelanders with a mean age of 76.6 years (table 1).

The IPF GWAS was a meta-analysis of three distinct cohorts (UK-, Colorado- and Chicago-based studies), which in total consisted of 2668 cases and 8591 controls [11]. The mean age was 67.3 years for cases and 64.7 years for controls. It is highly unlikely that there was any overlap of participants between the proteome and IPF GWASs, since they largely included

different geographical locations. Demographic characteristics from each study can be found in table 1 and the supplementary material.

### 4.5.2 Mendelian randomization

After MR scanning across 507 and 733 proteins from the two separate pQTL GWASs (834 total proteins, 406 of which were overlapped) for their association with IPF, three candidate proteins survived Benjamini–Hochberg correction: galactoside 3(4)-l-fucosyltransferase (FUT3), α-(1,3)-fucosyltransferase 5 (FUT5) and tumour necrosis factor receptor superfamily member 6B (TNFRSF6B) (table 2). FUT3 and FUT5 were replicated by the GWASs of both Sun *et al*. [12] and Emilsson *et al*. [13]. A 1 sd genetically determined higher plasma FUT3 and FUT5 had on average 19% and 24% lower risk of developing IPF (OR 0.81, 95% CI 0.74–0.88; $p=6.3\times10^{-7}$ and OR 0.76, 95% CI 0.68–0.86; $p=1.1\times10^{-5}$), respectively (table 2). Some previously described biomarkers for IPF, namely MMP1, MMP7 [6, 7] and CCL18 [9], and other members of the fucosyltransferase family (FUT8, FUT10 and POFUT1) were also assessed in this MR study. None showed causal effects on IPF risk (table 3, and supplementary tables S1 and S2). Supplementary tables S1 and S2 also show the results of all proteins analysed.

### 4.5.3 Colocalization analysis

We performed colocalisation analyses between the GWASs for candidate proteins (FUT3, FUT5 and TNFRSF6B) in Sun *et al*. [12] and the IPF GWAS to assess potential confounding due to LD. Both FUT3 and FUT5 were well colocalised with IPF by coloc with posterior probabilities of 99.9% and 97.7%, respectively, for a shared signal. TNFRSF6B had a lower posterior probability of 15.8% (figure 2). eCAVIAR estimated a high colocalisation joint posterior probability (CLPP) in FUT3 and FUT5 SNPs (0.28 and 0.016, respectively), but

TNFRSF6B had a low CLPP of $4.3\times10^{-6}$ (figure 2). Given the lack of clear colocalisation for TNFRSF6B, remaining analyses were focused on FUT3 and FUT5.

### 4.5.4 Sensitivity analyses

In Sun *et al*. [12], three *cis*-SNPs (rs104097772, rs12982233 and rs812936) were independently associated with FUT3 level when conditioned on the lead SNP (rs708686). One trans-SNP (rs679574) was also identified for FUT3 level. Two *cis*-SNPs (rs3760775 and rs4807054) were identified for FUT5, which were independently associated when conditioned on the lead SNP (rs778809). FUT3's trans-SNP (rs679574) was removed from analyses because it is palindromic and has a minor allele frequency of 0.49, making it impossible to harmonise with the IPF GWAS statistics. By using a method that can incorporate SNPs in LD [23], we included the other three *cis*-SNPs (rs104097772, rs12982233 and rs812936) that are in partial LD ($r^2\leq0.54$) with the sentinel SNP (rs708686). For FUT5, we included additional two *cis*-SNPs (rs3760775 and rs4807054) that are in partial LD ($r^2\leq0.12$) with the leading SNP (rs778809). The SNPs used were all identified in Sun *et al*. [12] and are listed in supplementary table S3. MR analyses, accounting for LD, using multiple *cis*-SNPs showed similar estimates both for FUT3 (IVW OR 0.84, 95% CI 0.78–0.91; $p=9.8\times10^{-6}$ and MR-Egger OR 0.69, 95% CI 0.50–0.97; p=0.03) and FUT5 (IVW OR 0.84, 95% CI 0.77–0.92; $p=1.4\times10^{-4}$ and MR-Egger OR 0.59, 95% CI 0.38–0.90; p=0.01) (table 4 and supplementary figure S1). The MR-Egger intercept estimates were close to the null, suggesting no detected evidence of directional pleiotropy (table 4). Bidirectional MR provided no evidence that IPF influences FUT3 and FUT5 levels (supplementary tables S4 and S5).

Although the FUT3/5 SNPs are on the same chromosome 19 as the genome-wide significant SNP in the IPF GWAS (rs12610495, near DPP9), they were not in LD (supplementary figure S2). However, given the LD between the FUT3 and FUT5 *cis*-SNPs (rs708686 and

rs778809/rs10420107; $r^2$=0.49), we performed statistical fine-mapping on the locus using FINEMAP [30] to explore the most important causal SNPs in the IPF GWAS [11]. The FUT3 SNP, rs708686, had the highest $\log_{10}$(Bayes factor (BF)) at 3.4 and the FUT5 SNPs, rs778809 and rs10420107, had a $\log_{10}$(BF) at 1.8, suggesting the FUT3 SNP had a higher probability of being causal for IPF (supplementary figure S3). Detailed methods are described in the supplementary material.

**Other shared genetic associations**

PhenoScanner searches identified that the FUT3 *cis*-SNP, rs708686, was also associated with an increased level of FUT5 [12] and decreased levels of vitamin B12 [31], lactoperoxidase [12], lithostathine-1-α [32] and FAM3B [12]. The FUT5 *cis*-SNPs, rs778809 and rs10420107, were associated with increased levels of FUT3 and decreased levels of FAM3B [12] (supplementary table S6). rs778809 was also associated with the plasma levels of CA19-9 and carcinoembryonic antigen (CEA) in individuals of Asian ancestry but the directions of the effects were not mentioned in the report [33]. Since we used *cis*-SNPs for FUT3 and FUT5, these pleiotropic effects on other molecules were more likely to represent vertical pleiotropy, where SNPs influencing levels of FUT3 and FUT5 in turn affect levels of the other molecules. Vertical pleiotropy does not violate the assumptions of MR. No other respiratory diseases or smoking habits were identified to be genome-wide significantly associated with the FUT3/5 *cis*-SNPs ($p<5\times10^{-8}$). We identified moderate associations between the FUT3 pQTL SNP and asthma (rs708686 allele T which decreases FUT3 level also decreases the risk of asthma; $p=1.1\times10^{-3}$) and between the FUT5 pQTL SNP and asthma (rs778809 allele A which decreases FUT5 level also decreases the risk of asthma; $p=3.4\times10^{-3}$) in the UK Biobank ($n_{cases}$=38 791).

Next, to reduce the possibility of biasing the MR estimates by horizontal pleiotropy of the FUT3/5 *cis*-SNPs, we performed MR to test if the aforementioned potential confounders, i.e. vitamin B12, lactoperoxidase, lithostathine-1-α, FAM3B, CA19-9 and CEA, could have an effect on IPF risk [34]. For these traits, only genetic determinants of each molecule identified in European ancestries were used. None of these potential confounders had evidence of their effects on IPF risk using MR (supplementary table S7). Figure 3 illustrates the overall findings. Detailed methods are described in the supplementary material.

**Literature search**

Further assessment for external validation of our findings involved a literature review by searching PubMed for reports published in English. The largest blood proteomic SOMAscan profiling study to date[35], involving 300 IPF patients and 100 matched controls for sex and smoking status, indicated that those with IPF had 0.89-fold lower level of FUT3 (log2FC: -0.18, p=0.019) but no difference in FUT5 level (log2FC: -0.024, p=0.76).

To assess the potential horizontal pleiotropy, we next searched for articles using the search terms “idiopathic pulmonary fibrosis” and each potential confounding factor, namely, vitamin B12, lactoperoxidase, lithostathine-1-alpha, FAM3B, CA19-9 and CEA. No previously published articles were found to describe the molecular mechanism of these factors in IPF pathophysiology.

### 4.5.5 Transcriptomic data of lung tissue

Using microarray-based transcriptomic data in whole lungs (GSE32537), we confirmed that a high *FUT3* expression level was associated with reduced risk of IPF (OR 0.50 per 1 sd increase, 95% CI 0.31–0.80; $p=3.4\times10^{-3}$), but *FUT5* was not clearly associated with IPF (OR

0.72 per 1 sd increase, 95% CI 0.46–1.1; p=0.14; $n_{case}/n_{control}$=119/50) (figure 4 and supplementary table S8).

scRNA-seq analyses from two public datasets (GSE135893 and GSE136831) revealed that *FUT3* was mainly expressed in epithelial cells in lungs (supplementary figure S5). There were distinct patterns of epithelial cell types between IPF and normal lung tissue. Alveolar type 2 cells were decreased and ciliated cells were increased in IPF lungs, which was in line with previous studies (supplementary figure S6) [36, 37]. *FUT3* expression in alveolar type 2 cells tended to be lower in IPF lungs than normal lungs ($p=1.9\times10^{-48}$ in GSE135893 and p=0.16 in GSE136831) (supplementary figure S7). Detailed results are described in the supplementary material.

### 4.6 Discussion

We undertook MR analyses of 834 circulating proteins to assess their effect on susceptibility to IPF in the largest GWASs of these traits available to date. Our analyses showed that subjects with genetically determined higher circulating levels of FUT3 and FUT5 had lower susceptibility to IPF. Colocalisation of FUT3/5 and IPF genetic signals and the absence of evidence of MR violations after thorough sensitivity analyses provided robust support for an aetiological effect of FUT3/5 on IPF susceptibility.

MR evidence for FUT3/5 was independently replicated using the GWASs of Sun *et al*. [12] and Emilsson *et al*. [13], which provide two distinct age distributions. Sun *et al*. [12] tested associations between protein levels and age, sex, BMI and estimated glomerular filtration rate (eGFR). They reported all proteins associated with either age, sex, BMI or eGFR with a significance threshold of $p<1\times10^{-5}$, whereby the positive association between age and FUT5

level ($p=1.6\times10^{-10}$) was described [12]. FUT3 level was not reported to be associated with any of the four demographic variables. In addition, neither FUT3 nor FUT5 was associated with age or sex among control samples (n=50) in publicly available bulk transcriptomic data in lungs (GSE32537). The genetic signals for IPF at the FUT3/5 locus were also consistent among three original IPF cohorts in the IPF GWAS study (supplementary table S9).

Given that the cost of measuring hundreds of proteins in adequately powered IPF studies involving samples collected years before disease onset is currently prohibitive, our approach provides an opportunity to prioritise candidate causal protein biomarkers by repurposing available data from large GWASs. MR studies for circulating biomarkers have often replicated or predicted the results of large-scale randomised controlled trials of pharmacological interventions to change biomarker levels [38–43]. Similarly, previous published biomarker studies have used the MR methodology to strengthen conclusions reported in the observational literature due to its robustness to reverse causation and most sources of confounding [44, 45]. Observational evidence sometimes provides opposite directions of effects to genetic findings, which is also the case for IPF. For example, rs207695 has been repeatedly shown to be associated with increased risk of IPF and the same variant is also known to decrease the expression of desmoplakin (DSP) in lungs and epithelial cells [11, 46, 47]. Taken together, this suggests that genetically low DSP expression leads to increased risk of IPF. On the other hand, some studies had identified that DSP is overexpressed in IPF lung tissue compared with normal lungs [46, 48], providing an opposite direction of effect. However, these observational results may be influenced by reverse causation, where IPF may influence the transcription of DSP. Nevertheless, an independent observational study demonstrated lower levels of circulating FUT3 in IPF patients [35] and our transcriptomic analyses also supported that increased FUT3 expression was associated with reduced risk of IPF.

It is still unclear how FUT3 may influence IPF risk. The fucosyltransferases encoded by FUT3 catalyse the formation of α-(1,4)-fucosylated glycoconjugates and are present only in two hominids (humans and chimpanzees). These genes are closely related, belonging to the Lewis FUT5–FUT3–FUT6 gene cluster, whose corresponding enzymes share 85% sequence similarity due to duplications of ancestral Lewis gene events [49]. Both FUT3 and FUT5 allow the synthesis of Lewis blood group antigens in exocrine secretions from precursor oligosaccharides [49]. Fucosylation is a post-translational modification that attaches fucose residues to polysaccharides, which partly determines mucin size and charge heterogeneity [50, 51]. PTS domain fucosylation in mucins could influence both the affinity to bind microorganisms and mucociliary clearance, consequently affecting the innate immune response and susceptibility to infections [52–54]. The gain-of-function mucin 5B (MUC5B) promoter SNP, rs35705950, has been repeatedly demonstrated to be associated with IPF risk [11, 55]. Overexpression of MUC5B in lungs was also shown to cause mucociliary dysfunction that enhances lung fibrosis in a mouse model [56]. These lines of evidence suggest a plausible link between MUC5B and fucosylation where host defences influence the pathophysiology of pulmonary fibrosis.

Elevated levels of CA19-9 had been shown to be associated with severity of pulmonary fibrosis [57]. However, our results found no evidence of this biomarker being causal for IPF. We observed that increased levels of FUT3 reduce susceptibility to IPF, which appears to contradict the previous studies since the FUT3 (Lewis) enzyme is known to be essential for biosynthesis of CA19-9 [58] and low levels of FUT3 lead to decreased levels of CA19-9. However, given that the pathology of IPF is characterised by microscopic honeycombing that is filled with mucus and inflammatory cells [59], this leads to overproduction of glycans, precursors of CA19-9. Concentrations of CA19-9 had been also noted to decline in IPF

patients after lung transplantation [60]. Elevated levels of CA19-9 are therefore likely to be a consequence of IPF.

Like all methods, our approach has important limitations. MR results may be biased by potential violations of its assumptions, which are not always confirmable, except for the SNP–exposure associations. However, our study design reduced potential horizontal pleiotropy by using *cis*-SNPs, which are backed by a biologically plausible rationale on protein levels and are unlikely to be mediated by other molecules. Furthermore, we undertook multiple sensitivity analyses to evaluate potential pleiotropic effects and did not identify evidence of horizontal pleiotropy for FUT3/5 and IPF. We also undertook colocalisation analyses, which additionally strengthened support for a shared genetic cause of FUT3/5 with IPF. Given the limited ethnicity of the current study population, further studies are needed to confirm the generalisability of these findings to non-European ancestry. Last, it was not ruled out in Sun *et al*. [12] that the association between *cis*-SNP rs708686 and FUT3 level measured by SOMAscan was influenced by potential epitope-binding artefacts driven by protein-altering variants. The negative MR findings of the causal relationships between established IPF biomarkers and IPF susceptibility could be attributed to the known evidence of modest correlations between some proteins measured by aptamer-based technology and those measured by immunoassay [61]. Such lack of correlation can lead to false-negative findings.

As the FUT3/5 pQTL SNPs were in LD and pleiotropic to each other, we could not confirm whether FUT3 and FUT5 had independent roles on IPF or whether they are influenced by each other. However, our sensitivity analyses and transcriptomic investigations suggested that FUT3 had a higher probability of being protective for IPF. There are no direct homologues of these proteins in mice and therefore in vivo functional follow-ups were not possible.

Alternatively, to test our results in a traditional observational study scenario, molar measurement of FUT3 in pre-diagnostic blood samples in larger, well-characterised, independent populations would be required. Unfortunately, at present, such samples are limited, given IPF's low incidence rate, but these should become more widely available with the development of large-scale population-based longitudinal biobanks.

In summary, undertaking an efficient MR scan of circulating proteins, our study demonstrated that genetically increased circulating FUT3 level is associated with reduced risk of IPF. These findings provide insights into the pathophysiology of this life-threatening disease, which may have potential translational relevance by identifying new targets for needed interventions.

## 4.7 Figures

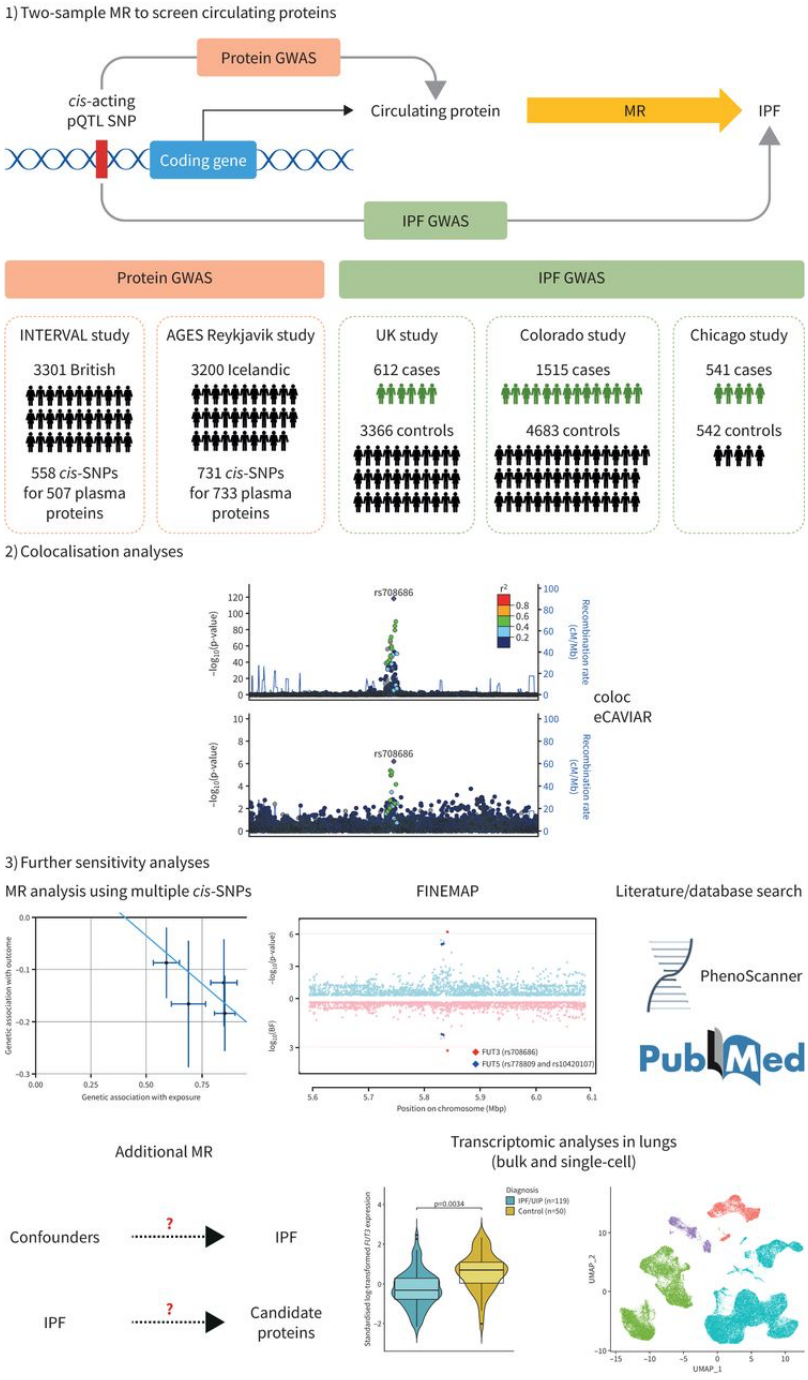


**Figure 1. Overall study design.**

See the main text and supplementary material for full details. MR: Mendelian randomisation; GWAS: genome-wide association study; pQTL: protein quantitative trait loci; SNP: single nucleotide polymorphism; IPF: idiopathic pulmonary fibrosis; UIP: usual interstitial pneumonia; UMAP: uniform manifold approximation and projection.

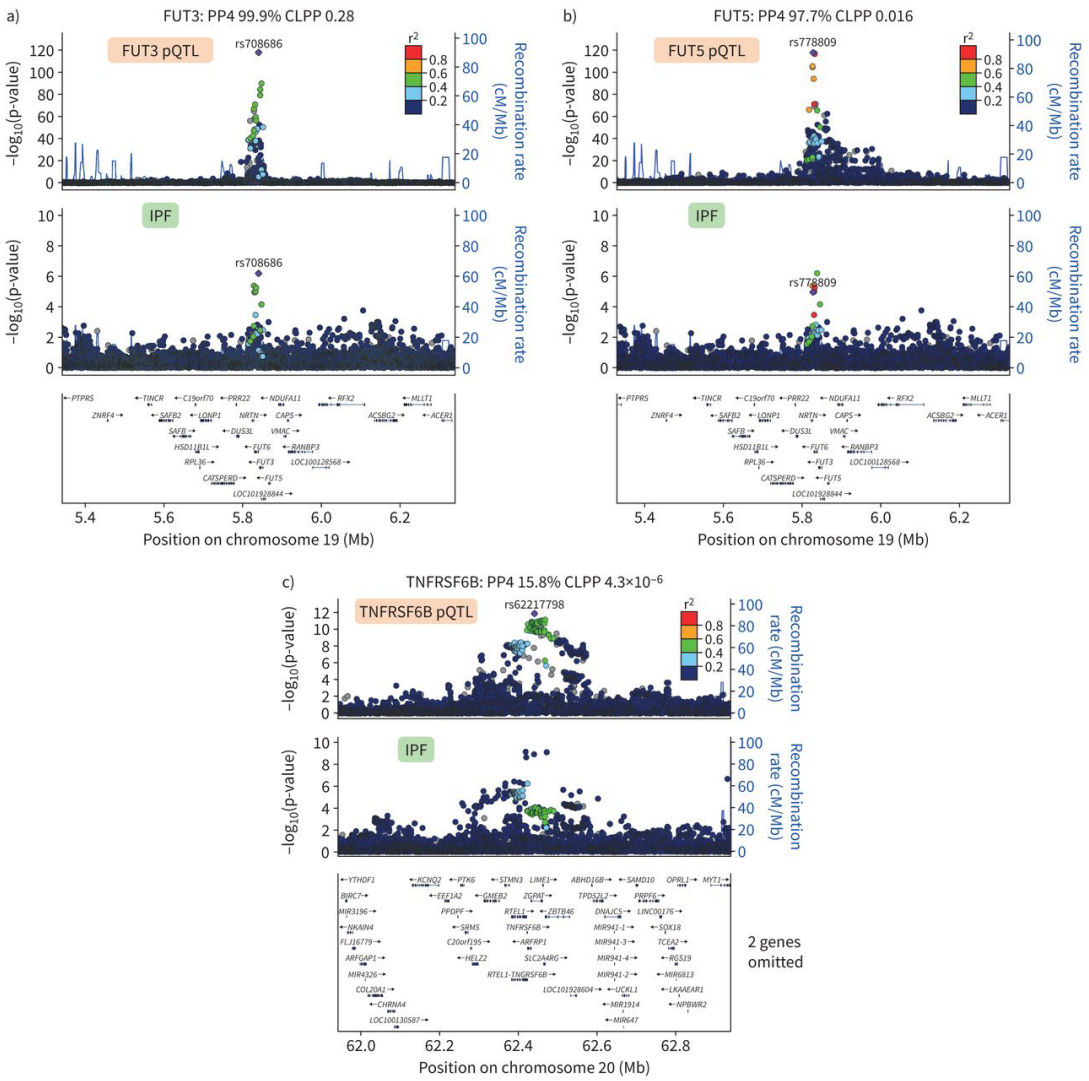


**Figure 2. Regional LocusZoom plots and the colocalization analyses results.**

Regional LocusZoom plots of three candidate idiopathic pulmonary fibrosis-influencing proteins: a) FUT3, b) FUT5 and c) TNFRSF6B. Each point represents a variant with chromosomal position on the x-axis (within 500-kb regions of each sentinel variant for candidate proteins) and the $-\log_{10}$(p-value) on the y-axis. Variants are coloured by linkage disequilibrium with the sentinel variant. Blue lines show the recombination rate; gene locations are shown at the bottom of the plot. PP4: posterior probability that the two traits share causal variants calculated by the coloc R package; CLPP: colocalisation joint posterior probability that the variants are causal for two traits calculated by eCAVIAR; pQTL: protein quantitative trait loci.

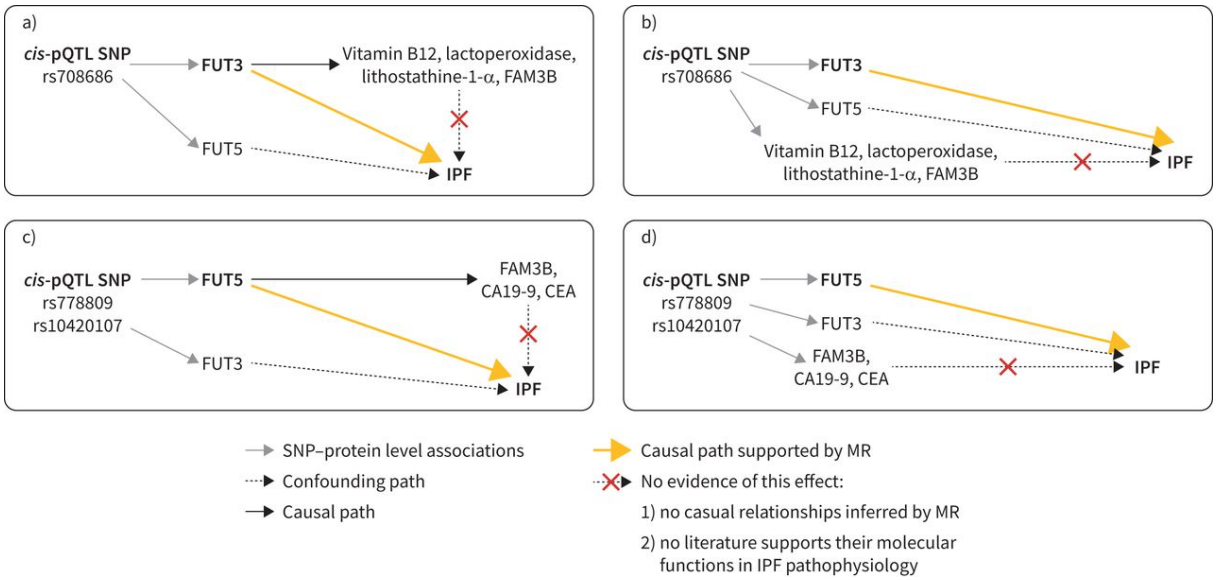


**Figure 3. Directed acyclic graphs illustrating the MR conclusions in four different scenarios.**

In all four scenarios, there was no evidence that the MR estimate of FUT3 and FUT5 on the idiopathic pulmonary fibrosis (IPF) risk was biased by violations of MR assumptions. Since we focused on *cis*-acting protein quantitative trait loci (pQTL) single nucleotide polymorphisms (SNPs) for FUT3 and FUT5, these pleiotropic effects on the levels of other molecules are more likely to be vertical pleiotropy rather than horizontal pleiotropy. Vertical pleiotropy occurs when *cis*-pQTL SNPs influence levels of FUT3 and FUT5 and these two proteins affect the levels of other molecules, which does not bias MR estimates. Moreover, in MR analysis using possible confounders as the exposure and IPF as the outcome, no causal relationships were validated. As FUT3/5 pQTL SNPs were in linkage disequilibrium and pleiotropic to each other, we could not confirm whether FUT3 and FUT5 had independent roles on IPF susceptibility. a) FUT3-associated *cis*-pQTL SNP rs708686 has an effect on IPF via FUT3 and FUT5. FUT3 has a direct effect on IPF and an indirect effect via vitamin B12, lactoperoxidase, lithostathine-1-α and FAM3B, which is an example of vertical pleiotropy that would not bias FUT3's MR estimate. However, this indirect effect was not supported by either MR evidence (supplementary table S7) or literature/database searches. b) FUT3-associated *cis*-pQTL SNP rs708686 has an effect on IPF via FUT3, FUT5 and potential

confounding variables: vitamin B12, lactoperoxidase, lithostathine-1-α and FAM3B. These confounders represent an example of horizontal pleiotropy that would bias FUT3's MR estimates. However, horizontal pleiotropic effects via these confounders were not supported by either MR analysis (supplementary table S7) or literature/database searches. c) FUT5-associated *cis*-pQTL SNPs rs778809 and rs10420107 have a direct effect on IPF via FUT5 and FUT3, and an indirect effect via FAM3B, CA19-9 and carcinoembryonic antigen (CEA). This indirect effect represents vertical pleiotropy and would not bias FUT5's MR estimate. However, this indirect effect was not supported by either MR evidence (supplementary table S7) or literature/database searches. d) FUT5-associated *cis*-pQTL SNPs rs778809 and rs10420107 have a direct effect on IPF via FUT5, FUT3 and potential confounding variables: FAM3B, CA19-9 and CEA. These confounders represent an example of horizontal pleiotropy that would bias FUT5's MR estimates. However, horizontal pleiotropic effects via these confounders were not supported by either MR analysis (supplementary table S7) or literature/database searches.

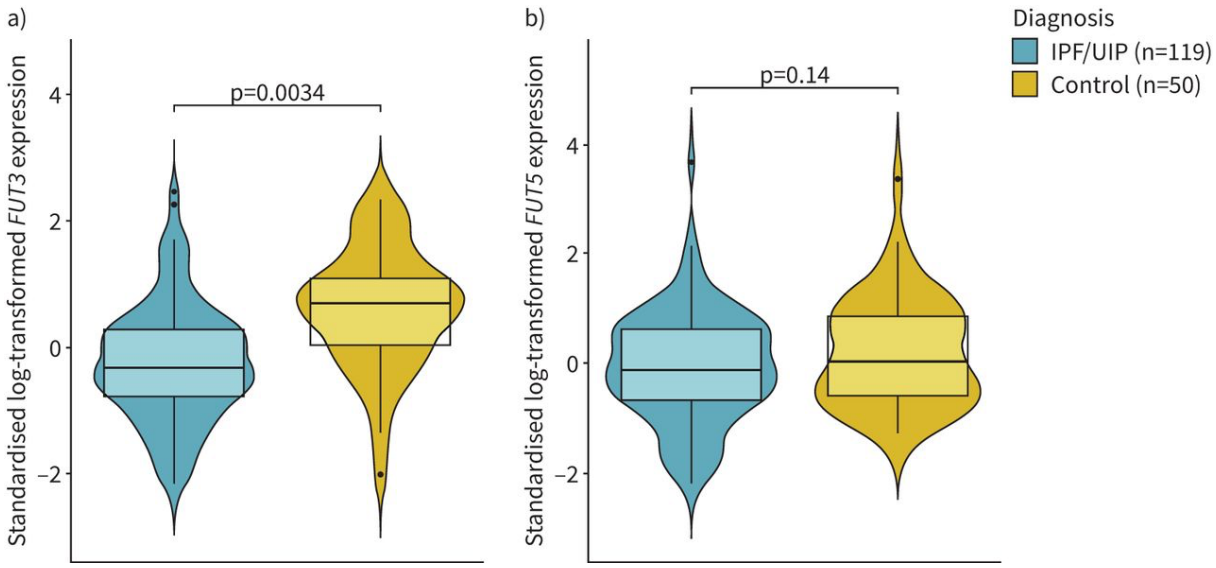


**Figure 4. a) FUT3 and b) FUT5 expression in whole lung compared between idiopathic pulmonary fibrosis (IPF)/usual interstitial pneumonia (UIP) and controls.**

This figure is based on data from microarray-based lung transcriptomic dataset GSE32537. Standardised log-transformed expression levels were compared between IPF/UIP (n=119) and controls (n=50). P-values were calculated by logistic regressions adjusted for age, sex and smoking status.

## 4.8 Tables

**Table 1. Demographic characteristics of the study cohorts.**

| | Sample size (n) | Ethnicity | Age (mean) (years) | Males (%) | Smokers (%) | Assay | Sample |
|---|---|---|---|---|---|---|---|
| **Proteome GWAS** | | | | | | | |
| Sun *et al.* [12] (INTERVAL study) | 3301 | British | 43.7 | 51.1 | 8.6[+] | SOMAscan | Plasma |
| Emilsson *et al.* [13] (AGES Reykjavik study) | 3200 | Icelandic | 76.6[#] | 42.7[#] | 12[#] | SOMAscan | Serum |
| **Allen *et al.* [11] (IPF GWAS)** | | | | | | | |
| Cases | 2668 | European | 67.3 | 69.3 | 72.5[§] | | |
| Controls | 8591 | European | 64.7[¶] | 57.1 | 66.1[§] | | |

GWAS: genome-wide association study; IPF: idiopathic pulmonary fibrosis. [#]: demographic characteristics were calculated with total participants in the AGES Reykjavik study (n=5457) (for smoking status, there was insufficient data to differentiate between current or ever-smokers); [¶]: mean age was calculated with samples from the Chicago- and UK-based studies (n=3908) since this information was not available for the Colorado-based study (supplementary material); [+]: percentage of current smokers; [§]: percentage of ever-smokers was calculated with samples from the Chicago- and UK-based studies (n=1153 for cases and n=3908 for controls) since this information was not available for the Colorado-based study (supplementary material).

**Table 2. Mendelian randomization (MR) analyses of the proteome for idiopathic pulmonary fibrosis.**

| | Chr. | Position (hg19) | SNP | Effect allele | Protein GWAS | | | | | IPF GWAS | | | MR estimate per increase in protein levels | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Protein | AF | Effect[#] | p-value | PVE (%) | AF | Effect | p-value | OR (95% CI) | p-value |
| Sun *et al.* [12] | 19 | 5840619 | rs708686 | C | FUT3 | 0.73 | 0.85 | $3.1\times10^{-273}$ | 27.3 | 0.72 | −0.18 | $6.3\times10^{-7}$ | 0.81 (0.74–0.88) | $6.3\times10^{-7}$ |
| (INTERVAL study) | 19 | 5830302 | rs778809 | G | FUT5 | 0.70 | 0.58 | $1.3\times10^{-118}$ | 14.0 | 0.68 | −0.16 | $1.1\times10^{-5}$ | 0.76 (0.68–0.86) | $1.1\times10^{-5}$ |
| Emilsson *et al.* [13] | 19 | 5840619 | rs708686 | C | FUT3 | 0.77 | 0.66 | $2.8\times10^{-126}$ | 21.0 | 0.72 | −0.18 | $6.3\times10^{-7}$ | 0.76 (0.68–0.84) | $6.3\times10^{-7}$ |
| (AGES Reykjavik study) | 19 | 5833279 | rs10420107 | G | FUT5 | 0.77 | 0.56 | $1.8\times10^{-91}$ | 11.7 | 0.68 | −0.16 | $9.2\times10^{-6}$ | 0.75 (0.66–0.85) | $9.2\times10^{-6}$ |
| | 20 | 62370349 | rs1056441 | T | TNFRSF6B | 0.39 | 0.14 | $2.0\times10^{-8}$ | 1.0 | 0.31 | −0.14 | $1.4\times10^{-4}$ | 0.38 (0.23–0.62) | $1.4\times10^{-4}$ |

Chr.: chromosome; SNP: single nucleotide polymorphism; GWAS: genome-wide association study; AF: allele frequency; PVE: phenotypic variance explained by the *cis*-protein quantitative trait loci SNP. [#]: in Sun *et al*. [12], each protein was first natural log-transformed and adjusted for age, sex, and duration between blood draw and processing, followed by rank-inverse normalisation; in Emilsson *et al*. [13], effect sizes were estimated for Yeo–Johnson-transformed protein level and thus we could not interpret the magnitude of the effect sizes.

**Table 3. Mendelian randomisation (MR) analyses of known idiopathic pulmonary fibrosis circulating biomarkers.**

| | Chr. | Position (hg19) | SNP | Effect allele | Protein GWAS | | | | | IPF GWAS | | | MR estimate per increase in protein levels | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Protein | AF | Effect[#] | p-value | PVE (%) | AF | Effect | p-value | OR (95% CI) | p-value |
| Emilsson *et al.* [13] (AGES Reykjavik study) | 11 | 102 697 731 | rs471994 | G | MMP1 | 0.66 | 0.55 | $7.0\times10^{-107}$ | 19.1 | 0.65 | −0.01 | 0.84 | 0.99 (0.87–1.12) | 0.84 |
| | 11 | 102 401 633 | rs11568819 | G | MMP7 | 0.95 | −0.50 | $5.0\times10^{-21}$ | 3.0 | 0.94 | −0.04 | 0.59 | 1.08 (0.82–1.42) | 0.59 |
| | 17 | 34 392 880 | rs712042 | T | CCL18 | 0.89 | −0.89 | $7.0\times10^{-124}$ | 13.4 | 0.86 | −0.04 | 0.42 | 1.05 (0.94–1.16) | 0.42 |

Chr.: chromosome; SNP: single nucleotide polymorphism; GWAS: genome-wide association study; AF: allele frequency; PVE: phenotypic variance explained by the *cis*-protein quantitative trait loci SNP. [#]: in Emilsson *et al*. [13], effect sizes were estimated for Yeo–Johnson-transformed protein level and thus we could not interpret the magnitude of the effect sizes.

**Table 4. Mendelian randomisation (MR) analyses considering linkage disequilibrium patterns using multiple *cis*-single nucleotide polymorphisms (SNPs) for FUT3 and FUT5.**

| Protein | Method | MR estimate per 1 SD increase in protein level | | Heterogeneity test | | Intercept | |
|---|---|---|---|---|---|---|---|
| | | OR (95% CI) | p-value | Test statistic | p-value | Intercept (95% CI) | p-value |
| FUT3 | IVW | 0.84 (0.78–0.91) | $9.8\times10^{-6}$ | 6.06 | 0.11 | | |
| | MR-Egger | 0.69 (0.50–0.97) | 0.03 | 3.98 | 0.14 | 0.15 (−0.09–0.38) | 0.23 |
| FUT5 | IVW | 0.84 (0.77–0.92) | $1.4\times10^{-4}$ | 7.19 | 0.03 | | |
| | MR-Egger | 0.59 (0.38–0.90) | 0.01 | 2.52 | 0.11 | 0.19 (−0.03–0.40) | 0.09 |

MR was performed using mr_inv and mr_egger functions in MendelianRandomisation version 0.4.3. Correlation matrices of SNPs were calculated using plink --r square with 503 individuals in the European subset of the 1000 Genomes Projects. We used a fixed effects inverse variance weighted (IVW) method and a random effects MR-Egger method.

## 4.9 List of references


1 Richeldi L, Collard HR, Jones MG. Idiopathic pulmonary fibrosis. *Lancet* 2017; 389: 1941–1952.

2 Vancheri C, Failla M, Crimi N, *et al*. Idiopathic pulmonary fibrosis: a disease with similarities and links to cancer biology. *Eur Respir J* 2010; 35: 496–504.

3 Richeldi L, Du Bois RM, Raghu G, *et al*. Efficacy and safety of nintedanib in idiopathic pulmonary fibrosis. *N Engl J Med* 2014; 370: 2071–2082.

4 King TE, Bradford WZ, Castro-Bernardini S, *et al*. A phase 3 trial of pirfenidone in patients with idiopathic pulmonary fibrosis. *N Engl J Med* 2014; 370: 2083–2092.

5 Hunninghake GM. Interstitial lung abnormalities: erecting fences in the path towards advanced pulmonary fibrosis. *Thorax* 2019; 74: 506–511.

6 Rosas IO, Richards TJ, Konishi K, *et al*. MMP1 and MMP7 as potential peripheral blood biomarkers in idiopathic pulmonary fibrosis. *PLoS Med* 2008; 5: e93.

7 White ES, Xia M, Murray S, *et al*. Plasma surfactant protein-D, matrix metalloproteinase-7, and osteopontin index distinguishes idiopathic pulmonary fibrosis from other idiopathic interstitial pneumonias. *Am J Respir Crit Care Med* 2016; 194: 1242–1251.

8 Kohno N, Kyoizumi S, Awaya Y, *et al*. New serum indicator of interstitial pneumonitis activity. Sialylated carbohydrate antigen KL-6. *Chest* 1989; 96: 68–73.

9 Neighbors M, Cabanski CR, Ramalingam TR, *et al*. Prognostic and predictive biomarkers for patients with idiopathic pulmonary fibrosis treated with pirfenidone: post-hoc assessment of the CAPACITY and ASCEND trials. *Lancet Respir Med* 2018; 6: 615–626.

10 Labrecque JA, Swanson SA. Interpretation and potential biases of Mendelian randomisation estimates with time-varying exposures. *Am J Epidemiol* 2019; 188: 231–238.

11 Allen RJ, Guillen-Guio B, Oldham JM, *et al*. Genome-wide association study of susceptibility to idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2020; 201: 564–574.

12 Sun BB, Maranville JC, Peters JE, *et al*. Genomic atlas of the human plasma proteome. *Nature* 2018; 558:73–79.

13 Emilsson V, Ilkov M, Lamb JR, *et al*. Co-regulatory networks of human serum proteins link genetics to disease. *Science* 2018; 361: 769–773.

14 Davey Smith G, Davies NM, Dimou N, *et al*. STROBE-MR: guidelines for strengthening the reporting of Mendelian randomisation studies. *Peer J Preprints* 2019; 7: 27857v.

15 Tam V, Patel N, Turcotte M, *et al*. Benefits and limitations of genome-wide association studies. *Nat Rev Genet* 2019; 20: 467–484.

16 Swanson SA, Hernan MA. The challenging interpretation of instrumental variable estimates under monotonicity. *Int J Epidemiol* 2018; 47: 1289–1297.

17 Davies NM, Holmes MV, Davey Smith G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *BMJ* 2018; 362: k601.

18 Hemani G, Zheng J, Elsworth B, *et al*. The MR-Base platform supports systematic causal inference across the human phenome. *Elife* 2018; 7: e34408.

19 Burgess S, Butterworth A, Thompson SG. Mendelian randomisation analysis with multiple genetic variants using summarized data. *Genet Epidemiol* 2013; 37: 658–665.

20 Giambartolomei C, Vukcevic D, Schadt EE, *et al*. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* 2014; 10: e100438.

21 Hormozdiari F, van de Bunt M, Segrè AV, *et al*. Colocalization of GWAS and eQTL signals detects target genes. *Am J Hum Genet* 2016; 99: 1245–1260.

22 Pruim RJ, Welch RP, Sanna S, *et al*. LocusZoom: regional visualisation of genome-wide association scan results. *Bioinformatics* 2010; 26: 2336–2337.

23 Yavorska OO, Burgess S. MendelianRandomisation: an R package for performing Mendelian randomisation analyses using summarized data. *Int J Epidemiol* 2017; 46: 1734–1739.

24 Burgess S, Dudbridge F, Thompson SG. Combining information on multiple instrumental variables in Mendelian randomisation: comparison of allele score and summarized data methods. *Stat Med* 2016; 35: 1880–1906.

25 Staley JR, Blackshaw J, Kamat MA, *et al*. PhenoScanner: a database of human genotype–phenotype associations. *Bioinformatics* 2016; 32: 3207–3209.

26 Kamat MA, Blackshaw JA, Young R, *et al*. PhenoScanner V2: an expanded tool for searching human genotype–phenotype associations. *Bioinformatics* 2019; 35: 4851–4853.

27 Yang IV, Coldren CD, Leach SM, *et al*. Expression of cilium-associated genes defines novel molecular subtypes of idiopathic pulmonary fibrosis. *Thorax* 2013; 68: 1114–1121.

28 Habermann AC, Gutierrez AJ, Bui LT, *et al*. Single-cell RNA sequencing reveals profibrotic roles of distinct epithelial and mesenchymal lineages in pulmonary fibrosis. *Sci Adv* 2020; 6: eaba1972.

29 Adams TS, Schupp JC, Poli S, *et al*. Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. *Sci Adv* 2020; 6: eaba1983.

30 Benner C, Spencer CCA, Havulinna AS, *et al*. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* 2016; 32: 1493–1501.

31 Nongmaithem SS, Joglekar CV, Krishnaveni GV, *et al*. GWAS identifies population-specific new regulatory variants in *FUT6* associated with plasma B12 concentrations in Indians. *Hum Mol Genet* 2017; 26: 2551–2564.

32 Yao C, Chen G, Song C, *et al*. Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nat Commun* 2018; 9: 3268.

33. He M, Wu C, Xu J, *et al*. A genome wide association study of genetic loci that influence tumour biomarkers cancer antigen 19-9, carcinoembryonic antigen and alpha fetoprotein and their associations with cancer risk. *Gut* 2014; 63: 143–151.

34. Hemani G, Tilling K, Davey Smith G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLoS Genet* 2017; 13: e1007081.

35. Todd JL, Neely ML, Overton R, *et al*. Peripheral blood proteomic profiling of idiopathic pulmonary fibrosis biomarkers in the multicentre IPF-PRO Registry. *Respir Res* 2019; 20: 227.

36. Parimon T, Yao C, Stripp BR, *et al*. Alveolar epithelial type II cells as drivers of lung fibrosis in idiopathic pulmonary fibrosis. *Int J Mol Sci* 2020; 21: 2269.

37. Plantier L, Crestani B, Wert SE, *et al*. Ectopic respiratory epithelial cell differentiation in bronchiolised distal airspaces in idiopathic pulmonary fibrosis. *Thorax* 2011; 66: 651–657.

38. Manousaki D, Mokry LE, Ross S, *et al*. Mendelian randomisation studies do not support a role for vitamin D in coronary artery disease. *Circ Cardiovasc Genet* 2016; 9: 349–356.

39. Manson JAE, Cook NR, Lee IM, *et al*. Vitamin D supplements and prevention of cancer and cardiovascular disease. *N Engl J Med* 2019; 380: 33–44.

40. Holmes MV, Smith GD. Dyslipidaemia: revealing the effect of CETP inhibition in cardiovascular disease. *Nat Rev Cardiol* 2017; 14: 635–636.

41. Holmes MV, Richardson TG, Ference BA, *et al*. Integrating genomics with biomarkers and therapeutic targets to invigorate cardiovascular drug development. *Nat Rev Cardiol* 2021; 18: 435–453.

42. Landray M. Tocilizumab in patients admitted to hospital with COVID-19 (RECOVERY): preliminary results of a randomised, controlled, open-label, platform trial. *Lancet* 2021; 397: 1637–1645.

43 Larsson SC, Burgess S, Gill D. Genetically proxied interleukin-6 receptor inhibition: opposing associations with COVID-19 and pneumonia. *Eur Respir J* 2021: 57: 2003545.

44. Fanidi A, Carreras-Torres R, Larose TL, *et al*. Is high vitamin B12 status a cause of lung cancer? *Int J Cancer* 2019; 145: 1499–1503.

45. Mokry LE, Ahmad O, Forgetta V, *et al*. Mendelian randomisation applied to drug development in cardiovascular disease: a review. *J Med Genet* 2015; 52: 71–79.

46. Mathai SK, Pedersen BS, Smith K, *et al*. Desmoplakin variants are associated with idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2016; 193: 1151–1160.

47. Moore C, Blumhagen RZ, Yang IV, *et al*. Resequencing study confirms that host defense and cell senescence gene variants contribute to the risk of idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2019; 200: 199–208.

48. Nance T, Smith KS, Anaya V, *et al*. Transcriptome analysis reveals differential splicing events in IPF lung tissue. *PLoS One* 2014; 9: e92111.

49. Dupuy F, Germot A, Marenda M, *et al*. α1,4-Fucosyltransferase activity: a significant function in the primate lineage has appeared twice independently. *Mol Biol Evol* 2002; 19: 815–824.

50. Johnson DC. Airway mucus function and dysfunction. *N Engl J Med* 2011; 364: 978.

51. Corfield AP. Mucins: a biologically relevant glycan barrier in mucosal protection. *Biochim Biophys Acta* 2015; 1850: 236–252.

52. Janssen WJ, Stefanski AL, Bochner BS, *et al*. Control of lung defence by mucins and macrophages: ancient defence mechanisms with modern functions. *Eur Respir J* 2016; 48: 1201–1214.

53. de Mattos LC. Structural diversity and biological importance of ABO, H, Lewis and secretor histo-blood group carbohydrates. *Rev Bras Hematol Hemoter* 2016; 38: 331–340.

54. Kerr SC, Fischer GJ, Sinha M, *et al*. *FleA* expression in Aspergillus fumigatus is recognized by fucosylated structures on mucins and macrophages to prevent lung infection. *PLoS Pathogens* 2016; 12: e1005555.

55. Seibold MA, Wise AL, Speer MC, *et al*. A common *MUC5B* promoter polymorphism and pulmonary fibrosis. *N Engl J Med* 2011; 364: 1503–1512.

56. Hancock LA, Hennessy CE, Solomon GM, *et al*. *Muc5b* overexpression causes mucociliary dysfunction and enhances lung fibrosis in mice. *Nat Commun* 2018; 9: 5363.

57. Maher TM, Oballa E, Simpson JK, *et al*. An epithelial biomarker signature for idiopathic pulmonary fibrosis: an analysis from the multicentre PROFILE cohort study. *Lancet Respir Med* 2017; 5: 946–955.

58. Kawai S, Suzuki K, Nishio K, *et al*. Smoking and serum CA19-9 levels according to Lewis and secretor genotypes. *Int J Cancer* 2008; 123: 2880–2884.

59. Raghu G, Collard HR, Egan JJ, *et al*. An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management. *Am J Respir Crit Care Med* 2011; 183: 788–824.

60 Rusanov V, Kramer MR, Raviv Y, *et al*. The significance of elevated tumor markers among patients with idiopathic pulmonary fibrosis before and after lung transplantation. *Chest* 2012; 141: 1047–1054.

61 Raffield LM, Dang H, Pratte KA, *et al*. Comparison of proteomic assessment methods in multiple cohort studies. *Proteomics* 2020; 20: 1900278.

## 4.10 Supplemental data

Supplementary Methods, Tables and Figures can be downloaded from the open access publication Nakanishi et al. in *Eur Respir J* available here: https://erj.ersjournals.com/content/59/2/2003979#sec-23

**Connecting Text: Bridge Between Chapter 4 and Chapter 5**

In the previous Chapter, we performed MR to identify the circulating protein with a potentially causal role in IPF susceptibility. We identified that plasma FUT3 is associated with a reduced risk of IPF, which could be an attractive therapeutic target for the disease.

Given the success in the previous Chapter, we showed that our MR approach for circulating proteins is a strong strategy to identify potentially druggable targets. Therefore, in the next Chapter, we applied the same strategy of MR with circulating proteins to the COVID-19 outcomes (severity, defined by critical illness [respiratory failure and/or death] and hospitalization, and susceptibility, defined by reported infection), to identify potentially druggable plasma proteins with etiologic role to COVID-19 outcomes. Despite the scale of the epidemic, few effective therapeutic options are available for the treatment of COVID-19. Thus, validated targets are needed for COVID-19 therapeutic development. Given the fact that MR studies in COVID-19 have already predicted the results of randomized controlled trials results, such as interleukin-6 inhibition(36), ACE inhibition(37), and vitamin D supplementation(38), the application of MR to COVID-19 may be a promising avenue to investigate potential opportunities for drug repurposing.

# Chapter 5: A Neanderthal OAS1 isoform protects individuals of European ancestry against COVID-19 susceptibility and severity

## 5.1 Title page

**A Neanderthal OAS1 isoform protects individuals of European ancestry against COVID-19 susceptibility and severity**

Sirui Zhou[1,2#], Guillaume Butler-Laporte[1,2#] , Tomoko Nakanishi[1,3,4,5#], David R. Morrison[1], Jonathan Afilalo[1,2,6], Marc Afilalo[1,7], Laetitia Laurent[1,] Maik Pietzner[8], Nicola Kerrison[8], Kaiqiong Zhao[1,2], Elsa Brunet-Ratnasingham[9,10], Danielle Henry[1], Nofar Kimchi[1], Zaman Afrasiabi[1], Nardin Rezk[1], Meriem Bouab[1], Louis Petitjean[1], Charlotte Guzman[1], Xiaoqing Xue[1], Chris Tselios[1], Branka Vulesevic[1], Olumide Adeleye[1], Tala Abdullah[1], Noor Almamlouk[1], Yiheng Chen[1,3], Michaël Chassé[9], Madeleine Durand[9], Clare Paterson[11], Johan Normark[12], Robert Frithiof[13], Miklós Lipcsey[13,14], Michael Hultström[13,15], Celia M. T. Greenwood [1,2,16], Hugo Zeberg[17], Claudia Langenberg[8,18], Elin Thysell[19] , Michael Pollak[1,20], Vincent Mooser[3], Vincenzo Forgetta[1], Daniel E. Kaufmann [9,21] and J. Brent Richards[1,2,3,22]

**[#]SZ, GBL and TN contributed equally to this work.**

**Affiliations:** [1]Lady Davis Institute, Jewish General Hospital, McGill University, Montréal, Quebec, Canada. [2]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Quebec, Canada. [3]Department of Human Genetics, McGill University, Montréal, Quebec, Canada. [4]Kyoto-McGill International Collaborative School in Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan. [5]Research Fellow, Japan Society for the Promotion of Science, Tokyo, Japan. [6]Department of Medicine, Division of Cardiology, McGill University, Montréal, Quebec, Canada. [7]Department of Emergency Medicine, Jewish General Hospital, McGill University, Montréal, Quebec, Canada. [8]MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Cambridge, UK. [9]Research Centre of the Centre Hospitalier de l'Université de

Montréal, Montréal, Quebec, Canada. [10]Department of Microbiology, Infectiology and Immunology, Université de Montréal, Montréal, Quebec, Canada. [11]SomaLogic, Inc., Boulder, CO, USA. [12]Molecular Infection Medicine Sweden (MIMS) and Wallenberg Center for Molecular Medicine, Department of Clinical Microbiology, Section of Infection and Immunology, Umeå University, Umeå, Sweden. [13]Anaesthesiology and Intensive Care Medicine, Department of Surgical Sciences, Uppsala University, Uppsala, Sweden. [14]Hedenstierna Laboratory, CIRRUS, Anaesthesiology and Intensive Care Medicine, Department of Surgical Sciences, Uppsala University, Uppsala, Sweden. [15]Integrative Physiology, Department of Medical Cell Biology, Uppsala University, Uppsala, Sweden. [16]Gerald Bronfman Department of Oncology, McGill University, Montréal, Quebec, Canada. [17]Department of Neuroscience, Karolinska Institutet, Stockholm, Sweden. [18]Computational Medicine, Berlin Institute of Health, Charité University Medicine Berlin, Berlin, Germany. [19]Department of Medical Biosciences, Pathology, Umeå University, Umeå, Sweden. [20]Departments of Medicine and Oncology, McGill University, Montréal, Quebec, Canada. [21]Department of Medicine, Université de Montréal, Montréal, Quebec, Canada. [22]Department of Twin Research, King's College London, London, UK.

**Published in:** *Nat Med* 27, 659–667 (2021). doi : 10.1038/s41591-021-01281-1

## 5.2 Abstract

To identify circulating proteins influencing Coronavirus Disease 2019 (COVID-19) susceptibility and severity, we undertook a two-sample Mendelian randomization (MR) study, rapidly scanning hundreds of circulating proteins while reducing bias due to reverse causation and confounding. In up to 14,134 cases and 1.2 million controls, we found that an s.d. increase in OAS1 levels was associated with reduced COVID-19 death or ventilation (odds ratio (OR) = 0.54, $P = 7 \times 10^{-8}$), hospitalization (OR = 0.61, $P = 8 \times 10^{-8}$) and susceptibility (OR = 0.78, $P = 8 \times 10^{-6}$). Measuring OAS1 levels in 504 individuals, we found that higher plasma OAS1 levels in a non-infectious state were associated with reduced COVID-19 susceptibility and severity. Further analyses suggested that a Neanderthal isoform of OAS1 in individuals of European ancestry affords this protection. Thus, evidence from MR and a case–control study support a protective role for OAS1 in COVID-19 adverse outcomes. Available pharmacological agents that increase OAS1 levels could be prioritized for drug development.

## 5.3 Introduction

To date, the COVID-19 pandemic has caused more than 2 million deaths worldwide and infected approximately 100 million individuals[1]. Despite the scale of the epidemic, there are, at present, few disease-specific therapies[2] to reduce the morbidity and mortality of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection. Apart from dexamethasone therapy in oxygen-dependent patients[3], most clinical trials have shown, at most, mild or inconsistent benefits on disease outcomes[4,5,6]. Therefore, validated targets are needed for COVID-19 therapeutic development.

One source of such targets is circulating proteins. Recent advances in large-scale proteomics have enabled the measurement of thousands of circulating proteins—and when combined with evidence from human genetics, such targets greatly improve the probability of drug development success[7,8,9]. Although de novo drug development will take time, the repurposing of currently available molecules targeting those proteins could provide an accelerated opportunity to deliver new therapies to patients.

Nevertheless, because confounding and reverse causation often bias traditional circulating protein studies, methods are needed to dissect causal relationships. This is especially the case in COVID-19, where exposure to SARS-CoV-2 unleashes profound changes in circulating protein levels[10]. One way to address these limitations is by using MR, a genetic epidemiology method that uses genetic variants as instrumental variables to test the effect of an exposure (here, protein levels) on an outcome (here, COVID-19 outcomes). The process of random assignment of alleles at conception greatly reduces bias from confounding. Because genotypes are always assigned before disease onset, MR studies are not influenced by reverse causation. However, MR rests on several assumptions[11], the most problematic being horizontal pleiotropy of the genetic instruments (wherein the genotype influences the

outcome, independently of the exposure). One way to help avoid this bias is to use genetic variants that influence circulating protein levels that are adjacent to the gene that encodes the circulating protein through the use of *cis*-protein quantitative trait loci (*cis*-pQTLs)[9]. *cis*-pQTLs are likely to influence the level of the circulating protein by directly influencing its transcription or translation and, therefore, less likely to affect the outcome of interest through pleiotropic pathways. Nevertheless, a causal genetic association between the exposure and outcome might be confounded by linkage disequilibrium (LD)[12], which can be detected through co-localization testing.

Understanding the etiologic role of circulating proteins in infectious diseases is challenging because the infection itself often leads to large changes in circulating protein levels[10]. Thus, it might appear that an increase in a circulating protein, such as a cytokine, is associated with a worsened outcome, when, in fact, the cytokine might be the host's response to this infection and help to mitigate this outcome. It is, therefore, important to identify genetic determinants of the protein levels in the non-infected state, which would reflect a person's baseline predisposition to the level of a protein.

MR studies can be complemented by traditional case–control studies, where the protein is longitudinally measured in patients with COVID-19 and controls, allowing for an estimation of the association between the protein level and COVID-19 outcomes. However, MR studies tend to predict the effect of the protein in the non-infectious state when the genetic determinants of such proteins are measured in the non-infected population. Because MR and case–control studies rely on different assumptions and might be influenced by different biases, concordant results between the two study designs can strengthen the cumulative evidence[13].

In this study, we, therefore, undertook two-sample MR and co-localization analyses to combine results from large-scale genome-wide association studies (GWASs) of circulating protein levels and COVID-19 outcomes[14]. We began by identifying the genetic determinants of circulating protein levels in large-scale proteomic GWASs and then used MR to assess whether these *cis*-pQTLs were associated with COVID-19 outcomes in large COVID-19 GWASs. Next, we investigated expression QTL (eQTL) and splice QTL (sQTL) effects of lead proteins. We then measured the most promising protein, OAS1, in individuals ascertained for SARS-CoV-2 infection, followed for longitudinal sampling during and after their infection.

## 5.4 Results

### 5.4.1 MR using *cis*-pQTLs and pleiotropy assessment.

The study design is illustrated in Fig. 1. We began by obtaining the genetic determinants of circulating protein levels from six large proteomic GWASs of individuals of European ancestry (Sun *et al.*[15] n = 3,301; Emilsson *et al.*[16] n = 3,200; Pietzner *et al.*[17] n = 10,708; Folkersen *et al.*[18] n = 3,394; Yao *et al.*[19] n = 6,861 and Suhre *et al.*[20] n = 997). A total of 931 proteins from these six studies had genome-wide significant *cis*-pQTLs or highly correlated LD proxies ($r^2 > 0.8$) in the meta-analyses of data from the COVID-19 Host Genetics Initiative[21], which included results from the GenOMICC program[22]. We then undertook MR analyses using 1,425 *cis*-pQTLs and 39 LD proxies as genetic instruments for circulating proteins in three COVID-19 outcomes: 1) very severe COVID-19 disease (defined as individuals experiencing death, mechanical ventilation, non-invasive ventilation, high-flow oxygen or use of extra-corporeal membrane oxygenation; 99.7% of these individuals were of European ancestry) using 4,336 cases and 623,902 controls; 2) COVID-19 disease requiring hospitalization using 6,406 cases and 902,088 controls of European ancestry; and 3) COVID-

19 susceptibility using 14,134 cases and 1,284,876 controls of European ancestry. In all outcomes, cases required evidence of SARS-CoV-2 infection. For the very severe COVID-19 and hospitalization outcomes, COVID-19 cases were defined as laboratory-confirmed SARS-CoV-2 infection based on nucleic acid amplification or serology tests. For the COVID-19 susceptibility outcome, cases were also identified by review of health records (using International Classification of Disease (ICD) codes or physician notes).

MR analyses revealed that the levels of three circulating proteins—2′–5′ oligoadenylate synthetase 1 (OAS1), interleukin-10 receptor beta subunit (IL10RB) and ABO—were associated with at least two COVID-19 outcomes after Benjamini–Hochberg false discovery rate correction (Table 1 and Supplementary Tables 1–6). Notably, increased OAS1 levels were strongly associated with protection from all three COVID-19 outcomes. Furthermore, these effect sizes were more pronounced with more severe outcomes, such that each s.d. increase in OAS1 levels was associated with decreased odds of very severe COVID-19 (OR = 0.54, 95% confidence interval (CI) 0.44–0.68, $P = 7.0 \times 10^{-8}$), hospitalization (OR = 0.61, 95% CI 0.51–0.73, $P = 8.3 \times 10^{-8}$) and susceptibility (OR = 0.78, 95% CI 0.69–0.87, $P = 7.6 \times 10^{-6}$) (Fig. 2a). We also identified OAS1 *cis*-pQTLs in Emilsson *et al.*[16] and Pietzner *et al.*[17], which were not included in the initial MR due to lack of genome-wide significance for their association with OAS1 levels16 or not included in their COVID-19 discovery panel[17]. MR analyses of using these additional *cis*-pQTLs yielded concordant results (Supplementary Table 7).

We next assessed whether the *cis*-pQTL for OAS1 levels (rs4767027) was associated with over 5,000 other diseases, traits or protein levels, as catalogued in PhenoScanner[23]. rs4767027 was not associated with any other traits or protein levels ($P < 5.0 \times 10^{-5}$). These findings reduce the possibility that the MR estimate of the effect of OAS1 on COVID-19 outcomes is

due to horizontal pleiotropy. Finally, except for COVID-19 susceptibility, the effect of rs4767027 did not demonstrate evidence of heterogeneity across COVID-19 Host Genetics Initiative GWAS meta-analyses (Table 1).

Using a *cis*-pQTL for IL10RB (rs2834167), we found that a 1-s.d. increase in circulating IL10RB level was associated with decreased odds of very severe COVID-19 (OR = 0.47, 95% CI 0.32–0.68, $P = 7.1 \times 10^{-5}$) and hospitalization (OR = 0.53, 95% CI 0.39–0.73, $P = 8.8 \times 10^{-5}$) but not susceptibility (Fig. 2a). Using PhenoScanner, we did not find evidence of pleiotropic effects of the c*is*-pQTL for IL10RB. A 1-s.d. increase in circulating ABO level was associated with increased odds of adverse COVID-19 outcomes (Table 1); however, we found that the *cis*-pQTL for ABO (rs505922) was strongly associated with the levels of several other proteins, suggesting potential horizontal pleiotropic effects (Supplementary Table 8). Given ABO's known involvement in multiple physiological processes, these results were expected but highlight that MR analyses might suffer from significant bias from horizontal pleiotropy.

### 5.4.2 Co-localization studies

To test whether confounding due to LD might have influenced the estimated effect of circulating OAS1 on COVID-19 outcomes, we tested the probability that the genetic determinants of OAS1 circulating protein level were shared with the three COVID-19 outcomes using co-localization analyses, as implemented in coloc[12]. The posterior probability that OAS1 levels and COVID-19 outcomes shared a single causal signal in the 1-Mb locus around the *cis*-pQTL, rs4767027, was 0.72 for very severe COVID-19, 0.82 for hospitalization due to COVID-19 and 0.89 for COVID-19 susceptibility (Fig. 3). This co-localization result was also replicated using OAS1 *cis*-pQTL identified by Pietzner *et al*.[17]

(Supplementary Table 7). This suggests that there is likely a single shared causal signal for OAS1 circulating protein levels and COVID-19 outcomes.

Co-localization of ABO levels and different COVID-19 outcomes also showed co-localization between ABO level and different COVID-19 outcomes (posterior probability of single shared signal = 0.90, 0.98 and 1 for ABO level and very severe COVID-19, hospitalization due to COVID-19 and susceptibility, respectively) (Extended Data Fig. 1). We were unable to perform co-localization analyses for IL10RB due to a lack of genome-wide summary-level data from the original proteomic GWAS[16].

### 5.4.3 Aptamer-binding effects

Protein-altering variants (PAVs)[15] might influence binding of affinity agents, such as aptamers or antibodies, that are used to quantify protein levels. We, thus, assessed if the *cis*-pQTLs for the MR-prioritized proteins were PAVs or in LD ($r^2 > 0.8$) with PAVs. rs2834167 (IL10RB) is a nonsense variant and could, therefore, be subject to potential binding effects. rs505922 (ABO) is not in LD with known missense variants. rs4767027 (OAS1) is an intronic variant, which is in LD with a missense variant rs2660 ($r^2 = 1$) in European ancestry. However, because expression studies derived from RNA sequencing are not subject to potential effects of missense variants that could influence aptamer binding, we next explored whether rs4767027 also influences OAS1 expression and/or splicing.

### 5.4.4 sQTL and eQTL studies for OAS genes

sQTLs are genetic variants that influence the transcription of different isoforms of a protein. The aptamer that targets OAS1 was developed against a synthetic protein comprising the amino acid sequence 1–364 of NP002525.2, which is common to the two major OAS1 isoforms: p46 and p42. Hence, the aptamer might identify both or either isoforms.

rs10774671 is a known sQTL for OAS1 that induces alternate splicing and creates p46 and p42 isoforms. Most present-day individuals of European ancestry carry the alternative variant (rs10774671-A). The ancestral variant (rs10774671-G) is the major allele in African populations and became fixed in Neanderthal and Denisovan genomes[24,25]. However, the ancestral variant, with its increased expression of the p46 isoform, was reintroduced into the European population via gene flow from Neanderthals[26]. Previous analyses suggest that individuals with either the GG or GA genotype at rs10774671 express higher amounts of p46 (ref. [26]), which is also the predominant isoform found in circulating blood[27]. Differences in antiviral activity have been observed between isoforms, with p46 being more active in certain viral infections[28]. Interestingly, the OAS1 pQTL rs4767027 is in high LD ($r^2 = 0.97$) with rs10774671 (ref. [26]) in European populations. Functional studies support that the G allele at rs10774671 increases expression of the p46 isoform but decreases expression of the p42 isoform[27]. This G allele at the sQTL rs10774671 reflects the T allele at pQTL rs4767027, which itself is associated with higher measured OAS1 levels and reduced odds of COVID-19 severity and susceptibility. These separate lines of evidence suggest that OAS1 levels, as measured by the SomaScan platform, predominantly identify the p46 isoform, which might protect against COVID-19 outcomes.

Undertaking MR studies of OAS1 splicing, we found that increased expression of the p46 isoform (as defined by normalized read counts of the intron cluster defined by LeafCutter[29,30]) was associated with reduced odds of COVID-19 outcomes (OR = 0.29, 95% CI 0.17–0.49, $P = 4.1 \times 10^{-6}$ for susceptibility, OR = 0.09, 95% CI 0.04–0.21, $P = 2.0 \times 10^{-8}$ for hospitalization and OR = 0.05, 95% CI 0.02–0.13, $P = 3.1 \times 10^{-9}$ for very severe COVID-19) (Fig. 2b). Co-localization analyses also supported a shared causal signal among the sQTL for OAS1, the pQTL and COVID-19 outcomes (Extended Data Fig. 2). Interestingly, the co-localization analyses supported a stronger probability of a shared signal with the sQTL than

the pQTL, suggesting that the p46 isoform might be the driver of the association of OAS1 levels with COVID-19 outcomes.

Next, we tested, using eQTL MR analyses, whether increased expression of OAS1 levels, without respect to isoform, was associated with COVID-19 outcomes. We identified an eQTL for total OAS1, rs10744785, from GTEx v8 (ref. [31]). Total OAS1 expression levels were not associated with COVID-19 susceptibility and hospitalization (Fig. 2b). We also found that increased OAS3 expression in whole blood was positively associated with COVID-19 outcomes in MR analyses and support for co-localization of their genetic signal (Extended Data Fig. 3 nd Supplementary Table 9).

Taken together, these pQTL, sQTL and eQTL studies suggest that increased levels of the p46 isoform of OAS1 seem to protect against COVID-19 adverse outcomes.

#### 5.4.5 Association of measured OAS1 protein level with COVID-19 outcomes.

Because MR studies were derived from protein levels measured in a non-infected state, we tested the hypothesis that increased OAS1 protein levels in a non-infected state would be associated with reduced odds of COVID-19 outcomes. To do so, we undertook a case–control study, measuring OAS1 protein levels using the SomaScan platform in 1,039 longitudinal samples from 399 patients who tested positive for SARS-CoV-2 by polymerase chain reaction (PCR) that were collected at multiple time points during their COVID-19 infection and 105 individuals who presented with COVID-19 symptoms but had negative SARS-CoV-2 PCR nasal swabs from the Biobanque Quebecoise de la COVID-19 cohort (www.BQC19.ca). Individuals who had undergone nasal swabs for SARS-CoV-2 infection were recruited prospectively (Table 2).

We defined non-infectious samples as those collected from convalescent patients with SARS-CoV-2 at least 31 d after onset of their symptoms (n = 115) or samples collected from patients negative for SARS-CoV-2 by PCR (n = 105). We also measured OAS1 levels in individuals with samples from patients positive for SARS-CoV-2 <14 d after symptom onset (n = 313), which showed increased OAS1 levels during infection (Extended Data Figs. 4–6). OAS1 levels are not associated with age and sex in non-infectious samples (Extended Data Fig. 7). After sample quality control (Methods), 308 patients with at least one sample collected during infection, 113 patients with at least one sample collected during a non-infectious state and 103 COVID-19-negative controls were included in the analyses (Extended Data Fig. 8).

To test whether OAS1 levels in a non-infectious state were associated with COVID-19 outcomes, we undertook logistic regression controlling for age, sex, age*age, plate, recruitment center and sample processing time. OAS1 levels were log-transformed and standardized to match the transformation procedure of the MR study. We found that, in the non-infectious samples, each s.d. increase in OAS1 levels on the log-transformed scale was associated with reduced odds of COVID-19 outcomes (OR = 0.20, 95% CI 0.08–0.53, P = 0.001 for very severe COVID-19; OR = 0.46, 95% CI 0.28–0.76, P = 0.002 for hospitalization; and OR = 0.69, 95% CI 0.49–0.98, P = 0.04 for susceptibility) (Fig. 4, Extended Data Fig. 9 and Supplementary Table 10). These results are consistent with our findings from MR, where increased circulating OAS1 levels in a non-infectious state were associated with protection against all of these adverse COVID-19 outcomes.

In samples drawn during active infection, we found that increased OAS1 levels were associated with increased odds of adverse COVID-19 outcomes (OR = 1.50, 95% CI 1.19–1.90, P = 0.0007 for very severe COVID-19; OR = 1.93, 95% CI 1.46–2.56, $P = 4.8 \times 10^{-6}$ for hospitalization; and OR = 4.39, 95% CI 2.87–6.73, $P = 1.09 \times 10^{-11}$ for susceptibility) (Fig. 4).

Taken together, these findings suggest that increased OAS1 levels in a non-infectious state are associated with better COVID-19 outcomes, and that, during infection, SARS-CoV-2 exposure likely causes OAS1 levels to increase, as interferon pathways are stimulated, which are known to increase OAS1 levels[32].

## 5.5 Discussion

Disease-specific therapies are needed to reduce the morbidity and mortality associated with COVID-19 outcomes. In this large-scale, two-sample MR study of 931 proteins assessed for three COVID-19 outcomes in up to 14,134 cases and 1.2 million controls of European ancestry, we provide evidence that increased OAS1 levels in the non-infectious state are strongly associated with reduced risks of very severe COVID-19, hospitalization and susceptibility. The protective effect size was particularly large, such that a 50% decrease in the odds of very severe COVID-19 was observed per s.d. increase in OAS1 circulating levels. OAS proteins are part of the innate immune response against RNA viruses. They are induced by interferons and activate latent RNase L, resulting in direct viral and endogenous RNA destruction, as demonstrated in in vitro studies[33]. Thus, OAS1 has a plausible biological activity against SARS-CoV-2. Because therapies exist that activate OAS1, repositioning them as potential COVID-19 treatments should be prioritized.

In populations outside of Sub-Saharan Africa, the protective alleles at both rs4767027-T (the OAS1 pQTL) and rs10774671-G (the OAS1 sQTL) are found on a Neanderthal haplotype[34], which was passed on to modern humans ~50,000–60,000 years ago[35]. The correspondence between the previously described gene flow[35] from Neandertals at this locus and the haplotype associated with protection against COVID-19 in the GWAS[22] was recently

demonstrated[34]. Even though these two single-nucleotide polymorphisms (SNPs) share a haplotype, their evolutionary histories differ. The rs4767027-T allele is derived from the Neanderthal lineage, whereas, for the rs10774671-G allele, Neanderthals preserved the ancestral state. OAS1 alternative splicing regulated by the rs10774671-G allele increases the isoform p46, which has a higher enzymatic activity against viruses than the p42 isoform[36] and is the only OAS1 isoform robustly upregulated during infection[26]. Although further studies are needed to fully elucidate the functional relevance of the pQTL and sQTL for OAS1, the antiviral activity of the gene products is higher for the Neandertal haplotype than the common haplotype in Europeans[28]. In Europeans, the Neandertal haplotype has undergone positive selection[26], and the rs4767027-T allele reaches an allele frequency of 0.32. Using MR and measurements of circulating proteins, we demonstrated here that increased OAS1 levels of the Neandertal haplotype in modern-day individuals of European ancestry confer this protective effect.

Our MR evidence indicated that higher p46 isoform levels of OAS1 and higher OAS1 total protein levels, as measured by the SomaScan assay, had protective effects on COVID-19 outcomes. These results were strongly supported by co-localization analysis. Given the consistent co-localization between the sQTL and pQTL for OAS1, the lack of co-localization between the eQTL and pQTL for OAS1 and the evidence that the SomaScan assay likely measures p46 isoforms, it seems probable that the protective effect of OAS1 is derived from the p46 isoform. However, further investigations are required to specifically measure each isoform in circulation, and isoform activity assays will be required to better understand if the p46 isoform, rather than total OAS1 levels, is most protective against COVID-19 outcomes.

The ancestral OAS1 splice variant encoding the more active p46 isoform was lost in the modern human population that left Africa. Several scenarios might explain this loss of

function—for example, loss of purifying selection during the out-of-Africa exodus, which might be due to changes in environmental pathogens or potential harm induced by OAS1 antiviral activity[37]. Unfortunately, we do not have sufficient data to test if the OAS1 p46 ancestral allele in Sub-Saharan Africans also offers protection against COVID-19. Nevertheless, these findings further emphasize the importance of the Neanderthal genome in COVID-19 risk modulation, because a risk locus on chromosome 3 has also been reported to be inherited from Neanderthals[38].

OAS1, OAS2 and OAS3 share considerable homology. As an interferon-stimulated gene[39], OAS1 polymorphisms have been associated with the host immune response to several classes of viral infection[40,41,42,43,44]. Given that OAS1 is an intracellular enzyme-activating RNase L leading to viral RNA degradation, it is probable that the circulating levels of this enzyme reflect intracellular levels of this protein. However, there is experimental evidence that extracellular OAS1 might also be important in the viral immune response[33].

Molecules currently exist that can influence OAS1 expression. Interferon beta-1b, which activates a cytokine cascade leading to increased OAS1 expression[45], is currently used to treat multiple sclerosis and has been shown to induce OAS1 expression in blood cells[46]. Interferon-based therapy has also been used in other viral infections[47]. However, recent randomized trials have shown inconsistent results. Although intravenous interferon beta-1b combined with lopinavir–ritonavir reduced mortality due to MERS-CoV infections[48], in the unblinded SOLIDARITY trial[49] there was no demonstrated benefit of intravenous interferon-beta-1b. On the other hand, a recent phase 2 trial testing the effect of inhaled nebulized interferon beta-1a (which is closely related to interferon beta-1b) showed improved COVID-19 symptoms in the treatment arm[50]. Although this study was not powered to show a difference in mortality, all deaths occurred in the placebo group. Inhaled nebulized interferon beta results in a much

higher tissue availability in the lung and might result in improved antiviral activity. Moreover, timing of administration is likely to play a role, as the administration of a pro-inflammatory cytokine might not provide benefit during the inflammation-driven phase of the disease. However, data on timing of administration are currently unavailable in the SOLIDARITY trial, and conclusions cannot yet be drawn. Lastly, the effect of interferon supplement might vary across ancestral populations, as different ancestries have different amounts of the more active p46 isoform of OAS1. Our study was limited to individuals of European ancestry, a population with higher expression of the p46 isoform. Interestingly, the SOLIDARITY trial enrolled 78% of its patients in South Asia, the Middle East, North Africa and Latin America, populations that might have higher expression of the p42 OAS1 isoform, whereas the study on inhaled interferon beta comprised 80% White patients from the United Kingdom. It is possible that interferon beta-1b might have different effects in populations of different ancestry due to different frequency of genetic variants in different populations.

There is in vitro evidence that pharmacological inhibition of phosphodiesterase-12, which degrades 2′–5′ oligoadenylate synthesized by OAS1, potentiates OAS-mediated antiviral activity[51,52]. Interestingly, coronaviruses in the same family as SARS-CoV-2 have been shown to produce viral proteins that degrade 2′–5′ oligoadenylate and reduce RNase-L activity, leading to evasion of the host immune response[53,54]. Our findings are also consistent with recent experimental work[55] showing that there are situations where SARS-CoV-2 is sensitive to OAS1-related antiviral defenses. Our findings motivate pharmacologic strategies to increase OAS1 levels or activity, as well as further evaluation of the possible antiviral activity of extracellular OAS1 (ref. [33]). Thus, existing preclinical molecules that lead to increased OAS1 levels[51] could be optimized and tested for their effect on COVID-19 outcomes.

Our MR analyses found that higher levels of OAS3 expression is associated with worse COVID-19 outcomes, which is an opposite direction of effect compared to OAS1. The discordant effects of the p46 isoform for OAS1 and OAS3 were also reported by a previous study[26], which might reflect complex biology of OAS genes for innate immune response. In a recent transcription-wide association study from the GenOMICC program[22], genetically predicted high expression of OAS3 in lungs and whole blood was associated with a higher risk of patients with COVID-19 becoming critically ill. Although further studies to assess the roles of OAS genes specific to SARS-CoV-2 are needed, it is likely that OAS1 is the main driver of the protective effect of the p46 isoform for COVID-19 outcomes given previous functional studies demonstrating the antiviral effect of OAS genes[26].

This study had limitations. First, we used MR to test the effect of circulating protein levels measured in a non-infected state because the effect of the *cis*-pQTLs on circulating proteins was estimated in individuals who had not been exposed to SARS-CoV-2. Once a person contracts SARS-CoV-2 infection, levels of circulating proteins could be altered, and this might be especially relevant for cytokines such as IL10 (which binds to IL10RB) and OAS1. Thus, the MR results presented in this paper should be interpreted as an estimation of the effect of circulating protein levels when measured in the non-infected state. Ongoing studies will help to clarify if the same *cis*-pQTLs influence circulating protein levels during infection. Second, this type of study suffers a high false-negative rate. Our goal was not to identify every circulating protein influencing COVID-19 outcomes but, rather, to provide evidence for a few proteins with strong *cis*-pQTLs, because these proteins are more likely to be robust to the assumptions of MR studies. Future large-scale proteomic studies with more circulating proteins properly assayed should help to overcome these limitations. Third, most MR studies assume a linear relationship between the exposure and the outcome. Thus, our findings would not identify proteins whose effect on COVID-19 outcomes has a clear threshold effect.

Fourth, the overall OAS1 levels measured by RNA sequencing (not only p46) might be biased by the effect of alternative splicing, and the role of overall OAS1 and OAS3 levels indicated by the association of the *cis*-pQTL of OAS1 in protection against COVID-19 are possible and not yet explored. We also could not completely exclude the possibility that measurement of OAS1 levels might be influenced by aptamer-binding effects. Last, all data presented in this paper pertain to individuals of European ancestry only—once again underlining the importance of genotyping efforts in other populations.

In conclusion, we used genetic determinants of circulating protein levels and COVID-19 outcomes obtained from large-scale studies and found compelling evidence that OAS1 has a protective effect on COVID-19 susceptibility and severity. Measuring plasma OAS1 levels in a case–control study demonstrated that higher circulating levels of this protein in a non-infectious state are strongly associated with reduced risk of adverse COVID-19 outcomes. Interestingly, the available evidence suggests that the protective effect from OAS1 in individuals of European ancestry is likely due to the Neanderthal-introgressed p46 OAS1 isoform. Known pharmacological agents that increase OAS1 levels[51] could be explored for their effect on COVID-19 outcomes.

## 5.6 Methods

### 5.6.1 pQTL GWAS

We systematically identified pQTL associations from six large proteomic GWASs[15,16,17,18,19,20]. Each of these studies undertook proteomic profiling using either SomaLogic SomaScans or O-link proximal extension assays.

### 5.6.2 COVID GWAS and COVID-19 outcomes

To assess the association of *cis*-pQTLs with COVID-19 outcomes, we used COVID-19 meta-analytic GWASs (data freeze 4) from the COVID-19 Host Genetics Initiative[21]. For our study, we used three of these GWAS meta-analyses, which included 25 cohorts of European ancestry and one cohort of admixed American ancestry. The outcomes tested were very severe COVID-19, hospitalization due to COVID-19 and susceptibility to COVID-19 (named A2, B2 and C2, respectively, by the COVID-19 Host Genetics Initiative).

Very severe COVID-19 cases were defined as hospitalized individuals with COVID-19 as the primary reason for hospital admission with laboratory-confirmed SARS-CoV-2 infection (nucleic acid amplification tests or serology based) and death or respiratory support (invasive ventilation, continuous positive airway pressure, bilevel positive airway pressure or continuous external negative pressure, high-flow nasal or face mask oxygen). Simple supplementary oxygen (for example, 2 L min−1 via nasal cannula) did not qualify for case status. Controls were all individuals in the participating cohorts who did not meet this case definition.

Hospitalized COVID-19 cases were defined as individuals hospitalized with laboratory-confirmed SARS-CoV-2 infection (using the same microbiology methods as for the very severe phenotype), where hospitalization was due to COVID-19-related symptoms. Controls were all individuals in the participating cohorts who did not meet this case definition.

Susceptibility to COVID-19 cases was defined as individuals with laboratory-confirmed SARS-CoV-2 infection, health record evidence of COVID-19 (ICD coding or physician confirmation) or with self-reported infections (for example, by questionnaire). Controls were all individuals who did not meet this case definition.

### 5.6.3 Two-sample MR

We used two-sample MR analyses to screen and test potential circulating proteins for their role in influencing COVID-19 outcomes. In two-sample MR, the effect of SNPs on the exposure and outcome are taken from separate GWASs. This method often improves statistical power because it allows for larger sample sizes for the exposure and outcome GWAS[56].

Exposure definitions: We conducted MR using six large proteomic GWAS studies[15,16,17,18,19,20]. Circulating proteins from Sun *et al*., Emilsson *et al*. and Pietzner *et al*. were measured on the SomaLogic platform; Suhre *et al*., Yao *et al*. and Folkersen *et al*. used protein measurements on the O-link platform. We selected proteins with only *cis*-pQTLs to test their effects on COVID-19 outcomes because they are less likely to be affected by potential horizontal pleiotropy. The *cis*-pQTLs were defined as the genome-wide significant SNPs ($P < 5 \times 10^{-8}$) with the lowest P value within 1 Mb of the transcription start site of the gene encoding the measured protein[9]. For proteins from Emilsson *et al*., Pietzner *et al*., Suhre *et al*., Yao *et al*. and Folkersen *et al*., we used the sentinel *cis*-pQTL per protein per study as these were the data available. For proteins from Sun *et al*., we used PLINK 1.9 (ref. [57]) and the 1000 Genome[58] European population reference panels to clump and select LD-independent *cis*-pQTL ($r^2 < 0.001$, distance 1,000 kb) with the lowest P value from reported summary statistics for each SOMAmer-bound protein. We included the same proteins represented by different *cis*-pQTLs from different studies to cross-examine the findings. For *cis*-pQTLs that were not present in the COVID-19 GWAS, SNPs with LD $r^2 > 0.8$ and with minor allele frequency (MAF) < 0.42 were selected as proxies; MAF > 0.3 was used for allelic alignment for proxy SNPs. *cis*-pQTLs with palindromic effects and with MAF > 0.42 were removed before MR to prevent allele mismatches. Benjamini–Hochberg correction was used

to control for the total number of proteins tested using MR. MR analyses were performed using the TwoSampleMR package in R[59]. For proteins with a single (sentinel) *cis*-pQTL, we used the Wald ratio to estimate the effect of each circulating protein on each of the three COVID-19 outcomes. For any proteins/SOMAmer reagents with multiple independent *cis*-pQTL, an inverse variance-weighted method was used to meta-analyze their combined effects. After harmonizing the *cis*-pQTLs of proteins with COVID-19 GWAS, a total of 566 SOMAmer reagents (529 proteins, 565 directly matched *cis*-pQTL and 26 proxies) from Sun *et al*., 760 proteins (747 directly matched *cis*-pQTL and 11 proxies) from Emilsson *et al*., 91 proteins (90 directly matched c*is*-pQTLs and two proxies) from Pietzner *et al*., 74 proteins (72 directly matched *cis*-pQTL) from Suhre *et al*., 24 proteins (24 directly matched *cis*-pQTLs) from Yao *et al*. and 13 proteins (13 directly matched *cis*-pQTLs) from Folkersen *et al*. were used as instruments for the MR analyses across the three COVID-19 outcomes (Supplementary Tables 11 and 12)[15,16,17,18,19,20].

### 5.6.4 Pleiotropy assessments

A common pitfall of MR is horizontal pleiotropy, which occurs when the genetic variant affects the outcome via pathways independent of the exposure. The use of circulating protein *cis*-pQTLs greatly reduces the possibility of pleiotropy, for reasons described above. We also searched in the PhenoScanner[23] database, a large catalog of observed SNP–outcome relationships involving >5,000 GWASs done to date to assess potentially pleiotropic effects of the *cis*-pQTLs of MR-prioritized proteins by testing the association of *cis*-pQTLs with other circulating proteins (that is, if they were trans-pQTLs to other proteins or significantly associated with other unrelated diseases or traits). For *cis*-pQTLs of MR-prioritized proteins measured on the SomaLogic platform, we assessed the possibility of potential aptamer-binding effects (where the presence of PAVs might affect protein measurements). We also

checked if *cis*-pQTLs of MR-prioritized proteins had significantly heterogeneous associations across COVID-19 populations in each COVID-19 outcome GWAS.

### 5.6.5 Co-localization analysis

Next, we tested co-localization of the genetic signal for the circulating protein and each of the three COVID-19 outcomes using co-localization analyses, which assess potential confounding by LD. Specifically, for each of these MR-significant proteins with genome-wide summary data available, for the proteomic GWASs a stringent Bayesian analysis was implemented in coloc[12] R package to analyze all variants in the 1-Mb genomic locus centered on the *cis*-pQTL. Co-localizations with posterior probability for hypothesis 4 (PP4, that there is an association for both protein level and COVID-19 outcomes, and they are driven by the same causal variant) > 0.5 were considered likely to co-localize (which means the highest posterior probability for all five coloc hypotheses), and PP4 > 0.8 was considered to be highly likely to co-localize.

### 5.6.6 sQTL and eQTL MR and co-localization studies for OAS genes

We performed MR and co-localization analysis using GTEx project v8 (ref. [31]) GWAS summary data to understand the effects of expression and alternative splicing of OAS genes in whole blood. The genetic instruments were conditionally independent ($r^2 < 0.001$) sQTLs and eQTLs for OAS1 and eQTLs for OAS2 and OAS3 identified by using stepwise regression in GTEx[31]. The sQTL SNP for OAS1 (rs10774671) was originally identified for the normalized read counts of LeafCutter[29] cluster of the last intron of the p46 isoform (chr12:112,917,700–112,919,389, GRCh38) in GTEx[30] and was used to estimate the effect of the p46 isoform. Co-localization analysis was performed using GWAS summary statistics from GTEx by restricting to the regions within 1 Mb of each QTL.

### 5.6.7 Measurement of plasma OAS1 protein levels associated with COVID-19 outcomes in BQC19

BQC19 is a Québec-wide initiative to enable research into the causes and consequences of COVID-19 disease. The patients included in this study were recruited at the Jewish General Hospital (JGH) and the Centre Hospitalier de l'Université de Montréal (CHUM) in Montréal, Québec, Canada.

COVID-19 case–control status was defined to be consistent with the GWAS study from the COVID-19 Host Genetics Initiative, from which the MR results were derived. Namely, we tested the association of OAS1 protein levels with the three different COVID-19 outcome definitions both in samples procured from non-infected stages and samples procured during the acute phase of the infection. The three outcomes were as follows. 1) Very severe COVID-19—defined as hospitalized individuals with laboratory-confirmed SARS-CoV-2 infection (nucleic acid amplification tests or serology based) and death or respiratory support (invasive ventilation, continuous positive airway pressure, bilevel positive airway pressure or continuous external negative pressure, high-flow nasal or face mask oxygen). Controls were all individuals who did not meet this case definition. 2) Hospitalized COVID-19 cases—defined as individuals hospitalized with laboratory-confirmed SARS-CoV-2 infection. Controls were all individuals who did not meet this case definition. 3) Susceptibility to COVID-19—cases were defined as individuals with laboratory-confirmed SARS-CoV-2 infection, and controls were all individuals who underwent PCR testing for SARS-CoV-2 but were negative. The date of symptom onset for patients with COVID-19 was collected from patient charts or estimated from their first positive COVID-19 tests if missing. Case inclusion criteria were not exclusive, which means that some individuals who were cases in the susceptibility analyses were also included in the hospitalization and very severe COVID-19 cohorts if they met case definitions.

A total of 125 individuals were recruited from CHUM, and 379 individuals were recruited from the JGH. Individuals had blood sampling done at up to five different time points (200 individuals had one measurement, 113 individuals had two measurements, 152 individuals had three measurements, 38 individuals had four measurements and one individual had five measurements). Days from symptom onset (T1) were calculated for each sample based on the date of symptom onset and blood draw date. For individuals who were negative for COVID-19, T1 was set to 0. Sample processing time (in hours) for each sample was also calculated to measure the duration of time from sample collection to processing to account for the increase in the amount of protein released from cell lysis due to extended sample handling time.

Protein levels in citrated (ACD) plasma samples were measured using the SomaScan assay. In total, 1,039 samples from 399 patients who were positive for SARS-CoV-2 and 105 patients who were negative for SARS-CoV-2 of mainly European descent underwent SomaScan assays, which included 5,284 SOMAmer reagents targeting 4,742 proteins. The SomaScan assay uses single-stranded DNA aptamers ('SOMAmers'), which are designed to selectively bind to a particular protein target[60]. SOMAmer reagent binding is quantified by microarray, measuring abundance in relative fluorescent units (RFUs). The RFUs for each protein underwent four normalization processes, including hybridization control, intraplate median signal normalization, plate scaling and calibration and median signal normalization to a reference generated from internal data across all samples. All normalizations were conducted by SomaLogic and detailed in their Technical Note[61].

Of particpants who were positive for SARS-CoV-2, we defined samples procured from patients during the infectious state as those sampled within 14 d (including the 14th day) from the first date of symptoms[62]. For patients with more than one sample within 14 d of symptom

onset, the earliest sample was used. We defined samples procured from patients who were non-infectious as samples from SARS-CoV-2-positive patients taken at least 31 d after symptom onset. We selected 31 d, as this is the upper limit of the interquartile range of the duration of SARS-CoV-2 positivity in a recent systematic review and coincided with the first scheduled outpatient follow-up blood test in the BQC19 (ref. [63]). For individuals with more than one sample at least 31 d after symptom onset, the latest sample was used.

OAS1 level was measured by one SOMAmer reagent (OAS1.10361.25). Within each group, median signal-normalized OAS1 levels were natural log-transformed and adjusted for sample processing time, and the residuals were further standardized. For each group, we removed samples that were outliers with long sample processing time (sample processing time > 50 h) or high OAS1 level (log OAS1 level > 8). Logistic regression was performed to test the association-standardized OAS1 level with the three COVID-19 outcomes including age, sex, age*age, center of recruitment and plates as covariates.

### 5.7 Data availability

Data from proteomics studies and GTEx consortium (GTEx project v8 (ref. [31])) are available from the referenced peer-reviewed studies[15,16,17,18,19,20] or their corresponding authors, as applicable. The PhenoScanner online database is available at http://www.phenoscanner.medschl.cam.ac.uk/. Summary statistics for the COVID-19 outcomes are publicly available for download on the COVID-19 Host Genetics Initiative website (www.covid19hg.org). The BQC19 is an Open Science biobank. Instructions on how to access data for individuals from the BQC19 at the Jewish General Hospital site are available here: https://www.mcgill.ca/genepi/mcg-covid-19-biobank. Instructions on how to access data from other sites of the BQC19 are available here: https://www.bqc19.ca/en/access-data-samples.

## 5.8 Figures

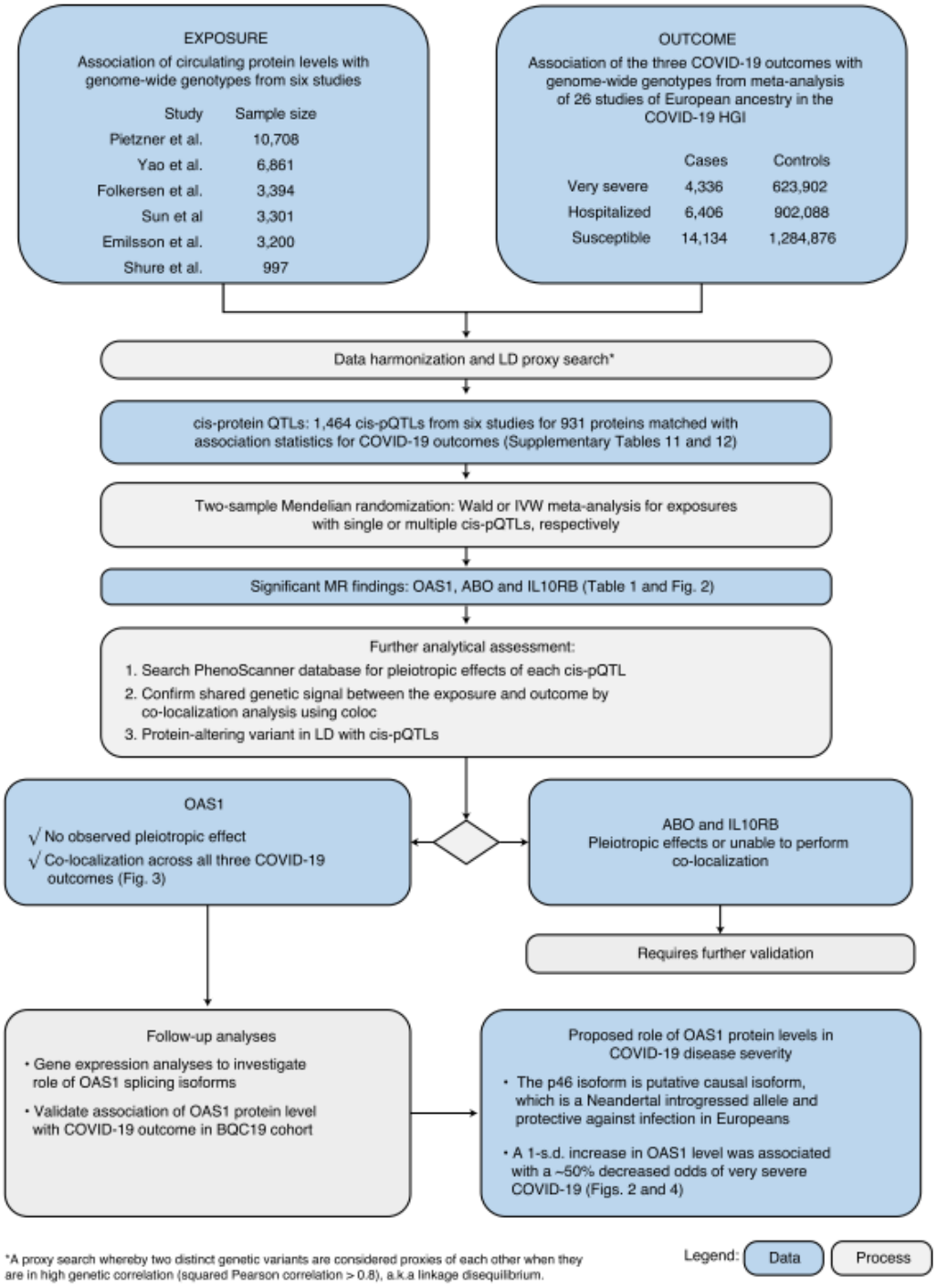


**Figure 1. Flow diagram of study design.**

IVW, inverse variance-weighted.

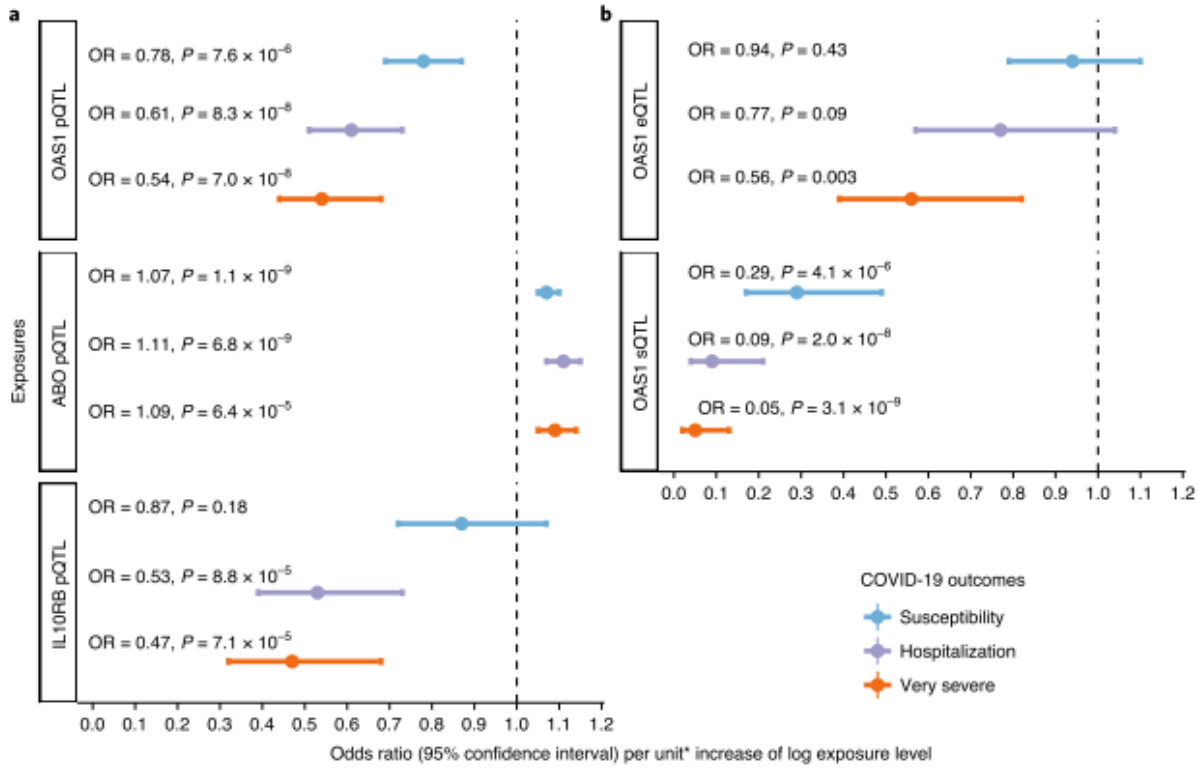


**Figure 2. Association of circulating protein levels of OAS1, ABO and IL10RB and messenger RNA levels of OAS1 with COVID-19 outcomes from MR.**

Forest plot showing OR and 95% CI from two sample MR analyses (two sided). P values are unadjusted. a, MR estimates of proteins influencing COVID- 19 outcomes; unit: s.d. of log-normalized value. b, MR estimates of OAS1 messenger RNA influencing COVID-19 outcomes; unit: s.d. of normalized read counts.

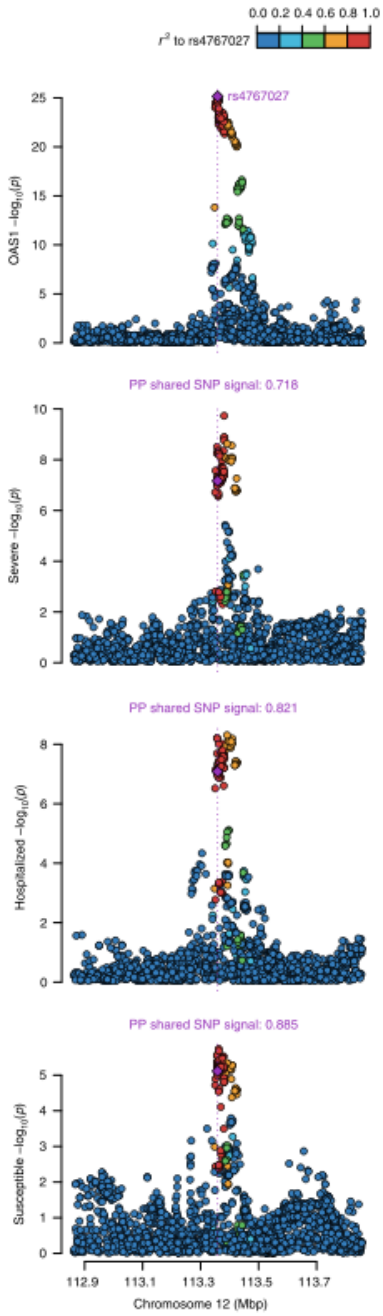


**Figure 3. Co-localization of the genetic determinants of OAS1 plasma protein levels and COVID-19 outcomes.**

Co-localization of genetic signal for OAS1 levels (top plot) and COVID-19 outcomes (three bottom plots) in the 1-Mb region around OAS1 pQTL rs4767027; color shows SNPs in the region in LD ($r^2$) with rs4767027 (purple). The posterior probability (PP) of a shared single signal between OAS1 levels and the three COVID-19 outcomes was estimated by coloc.

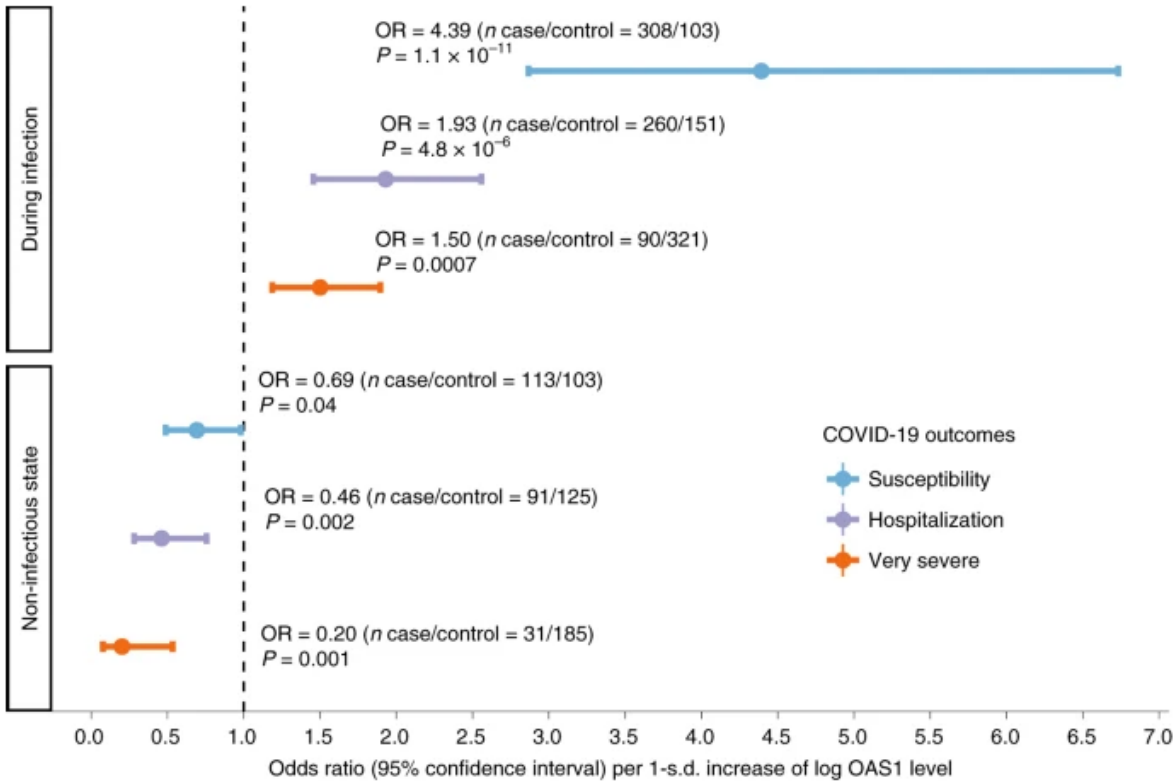


**Figure 4. Association of OAS1 levels with COVID-19 outcomes from the case–control study in BQC19.**

Forest plot showing ORs and 95% CIs from logistic regression analyses (two sided). P values are unadjusted. During infection: patient samples that were collected within 14 d from the date of symptom onset. For individuals with two or more samples collected within 14 d of symptom onset, the earliest time point was used. Non-infectious state: patient samples that were collected at least 31 d from the date of symptom onset. For individuals with two or more samples collected at different time points at least 31 d from symptom onset, the latest time point was used. Additional information is also described in Supplementary Table 10.

## 5.8 Tables

**Table 1. MR-identified circulating protein levels affecting COVID-19 outcomes.**

| Protein | cis-pQTL | Source | Very severe COVID-19 (99.7% European ancestry) | | | | COVID-19 hospitalization (European ancestry only) | | | | COVID-19 susceptibility (European ancestry only) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | OR | 95% CI | *P* value | *P* het | OR | 95% CI | *P* value | *P* het | OR | 95% CI | *P* value | *P* het |
| OAS1 | rs4767027 | Sun | 0.54 | 0.44–0.68 | $7.0\times10^{-8}$ | 0.37 | 0.61 | 0.51–0.73 | $8.3\times10^{-8}$ | 0.16 | 0.78 | 0.69–0.87 | $7.6\times10^{-6}$ | 0.005 |
| ABO | rs505922 | Sun, Emilsson | 1.09 | 1.05–1.14 | $6.4\times10^{-5}$ | 0.10 | 1.11 | 1.07–1.15 | $6.8\times10^{-9}$ | 0.06 | 1.07 | 1.05–1.10 | $1.1\times10^{-9}$ | 0.10 |
| IL10RB | rs2834167 | Emilsson | 0.47 | 0.32–0.68 | $7.1\times10^{-5}$ | 0.02 | 0.53 | 0.39–0.73 | $8.8\times10^{-5}$ | 0.11 | 0.87 | 0.72–1.07 | 0.18 | 0.006 |

OR represents the estimated effect of an s.d. on the natural log-scale (for Sun *et al*.) or one-unit (for Emilsson *et al*.) increase in protein levels on the odds of the three COVID-19 outcomes. P het, P value of heterogeneity for each *cis*-pQTL across the cohorts in the GWAS summary-level meta-analysis from the COVID-19 Host Genomic Initiative.

**Table 2. Participant demographics of the BQC19 cohort included in this study.**

| Sample demographics | Number of individuals (%) (total $n = 504$) |
|---|---|
| Sex | |
| Female | 250 (49.6%) |
| Male | 254 (50.4%) |
| Age (years) [a] | 65.4 (18.0) |
| Body mass index[a] | 28.6 (6.18) |
| Missing | 225 (44.6%) |
| SARS-CoV-2 PCR test | |
| Positive | 399 (79.2%) |
| Negative | 105 (20.8%) |
| Hospitalization | |
| Hospitalized | 406 (80.6%) |
| Outpatient treatment only | 98 (19.4%) |
| Hospitalization duration (d)[b] | 14.0 (6.00, 27.0) |
| Death | |
| Deceased | 43 (8.5%) |
| Survived | 461 (91.5%) |
| Respiratory support | |
| No oxygen | 233 (46.2%) |
| Oxygen supplementation | 143 (28.4%) |
| Mechanical ventilation | 128 (25.4%) |
| Days on ventilator[b] | 14.0 (6.75, 23.5) |

[a]Mean (s.d.) [b]Median (25% interquartile range and 75% interquartile range), which was calculated among individuals who were hospitalized and individuals on a ventilator, respectively.

## 5.9 List of references


1. Johns Hopkins University of Medicine. Coronavirus Resource Center. https://coronavirus.jhu.edu/ (2020).

2. Weinreich, D. M. *et al*. REGN-COV2, a neutralizing antibody cocktail, in outpatients with Covid-19. *N. Engl. J. Med*. 384, 238–251 (2020).

3. Horby, P. *et al*. Dexamethasone in hospitalized patients with Covid-19—preliminary report. *N. Engl. J. Med.* https://doi.org/10.1056/NEJMoa2021436 (2020).

4. Sterne, J. A. C. *et al*. Association between administration of systemic corticosteroids and mortality among critically ill patients with COVID-19: a meta-analysis. *JAMA* 324, 1330–1341 (2020).

5. Beigel, J. H., Tomashek, K. M. & Dodd, L. E. Remdesivir for the treatment of Covid-19—preliminary report. *N. Engl. J. Med*. 383, 994 (2020).

6. Cavalcanti, A. B. *et al*. Hydroxychloroquine with or without azithromycin in mild-to-moderate Covid-19. *N. Engl. J. Med*. 383, 2041–2052 (2020).

7. Nelson, M. R. *et al*. The support of human genetic evidence for approved drug indications. *Nat. Genet.* 47, 856–860 (2015).

8. Cook, D. *et al.* Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. *Nat. Rev. Drug Discov*. 13, 419–431 (2014).

9. Zheng, J. *et al.* Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nat. Genet*. 52, 1122–1131 (2020).

10. Filbin, M. R. *et al*. Plasma proteomics reveals tissue-specific cell death and mediators of cell–cell interactions in severe COVID-19 patients. Preprint at *bioRxiv* https://doi.org/10.1101/2020.11.02.365536 (2020).

11. Davey Smith, G., Ebrahim, S., Smith, G. D. & Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol*. 32, 1–22 (2003).

12. Giambartolomei, C. *et al*. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet*. 10, e1004383 (2014).

13. Lawlor, D. A., Tilling, K. & Smith, G. D. Triangulation in aetiological epidemiology. *Int. J. Epidemiol*. 45, 1866–1886 (2016).

14. COVID-19 Host Genetics Initiative. The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. *Eur. J. Hum. Genet*. 28, 715–718 (2020).

15. Sun, B. B. *et al*. Genomic atlas of the human plasma proteome. *Nature* 558, 73–79 (2018).

16. Emilsson, V. *et al*. Co-regulatory networks of human serum proteins link genetics to disease. *Science* 361, 769–773 (2018).

17. Pietzner, M. *et al*. Genetic architecture of host proteins involved in SARS-CoV-2 infection. *Nat. Commun*. 11, 6397 (2020).

18. Folkersen, L. *et al*. Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease. *PLoS Genet*. 13, e1006706 (2017).

19. Yao, C. *et al*. Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nat. Commun*. 9, 3268 (2018).

20. Suhre, K. *et al*. Connecting genetic risk to disease end points through the human blood plasma proteome. *Nat. Commun*. 8, 14357 (2017).

21. COVID-19 Host Genetics Initiative. https://www.covid19hg.org/results/ (2021).

22. Pairo-Castineira, E. *et al*. Genetic mechanisms of critical illness in Covid-19. *Nature* https://doi.org/10.1038/s41586-020-03065-y (2020).

23. Staley, J. R. *et al*. PhenoScanner: a database of human genotype–phenotype associations. *Bioinformatics* 32, 3207–3209 (2016).

24. Prüfer, K. *et al.* A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* 358, 655–658 (2017).

25. Meyer, M. *et al*. A high-coverage genome sequence from an archaic Denisovan individual. *Science* 338, 222–226 (2012).

26. Sams, A. J. *et al*. Adaptively introgressed Neandertal haplotype at the OAS locus functionally impacts innate immune responses in humans. *Genome Biol*. 17, 246 (2016).

27. Li, H. *et al*. Identification of a Sjögren's syndrome susceptibility locus at OAS1 that influences isoform switching, protein expression, and responsiveness to type I interferons. *PLoS Genet*. 13, e1006820 (2017).

28. Liu, X. *et al*. A functional variant in the OAS1 gene is associated with Sjögren's syndrome complicated with HBV infection. *Sci. Rep*. 7, 17571 (2017).

29. Li, Y. I. *et al*. Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet*. 50, 151–158 (2018).

30. Aguet, F. *et al.* The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 369, 1318–1330 (2020).

31. Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nature* 550, 204–213 (2017).

32. Hornung, V., Hartmann, R., Ablasser, A. & Hopfner, K. P. OAS proteins and cGAS: unifying concepts in sensing and responding to cytosolic nucleic acids. *Nat. Rev. Immunol.* 14, 521–528 (2014).

33. Kristiansen, H. *et al.* Extracellular 2′–5′ oligoadenylate synthetase stimulates RNase L-independent antiviral activity: a novel mechanism of virus-induced innate immunity. *J. Virol.* 84, 11898–11904 (2010).

34. Zeberg, H. & Pääbo, S. A genomic region associated with protection against severe COVID-19 is inherited from Neandertals. *Proc. Natl Acad. Sci. USA* 118, e2026309118 (2021).

35. Mendez, F. L., Watkins, J. C. & Hammer, M. F. Neandertal origin of genetic variation at the cluster of OAS immunity genes. *Mol. Biol. Evol.* 30, 798–801 (2013).

36. Bonnevie-Nielsen, V. *et al.* Variation in antiviral 2′–5′-oligoadenylate synthetase (2′5′AS) enzyme activity is controlled by a single-nucleotide polymorphism at a splice-acceptor site in the OAS1 gene. *Am. J. Hum. Genet*. 76, 623–633 (2005).

37. Carey, C. M. *et al*. Recurrent loss-of-function mutations reveal costs to OAS1 antiviral activity in primates. *Cell Host Microbe* 25, 336–343 (2019).

38. Zeberg, H. & Pääbo, S. The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. *Nature* 587, 610–612 (2020).

39. Schneider, W. M., Chevillotte, M. D. & Rice, C. M. Interferon-stimulated genes: a complex web of host defenses. *Annu. Rev. Immunol*. 32, 513–545 (2014).

40. Min, J.-Y. & Krug, R. M. The primary function of RNA binding by the influenza A virus NS1 protein in infected cells: inhibiting the 2′–5′ oligo (A) synthetase/RNase L pathway. *Proc. Natl Acad. Sci. USA* 103, 7100–7105 (2006).

41. Hu, B. *et al*. Cellular responses to HSV-1 infection are linked to specific types of alterations in the host transcriptome. *Sci. Rep*. 6, 28075 (2016).

42. Lim, J. K. *et al*. Genetic variation in OAS1 is a risk factor for initial infection with West Nile virus in man. *PLoS Pathog*. 5, e1000321 (2009).

43. Simon-Loriere, E. *et al*. High anti-dengue virus activity of the OAS gene family is associated with increased severity of dengue. *J. Infect. Dis*. 212, 2011–2020 (2015).

44. Hamano, E. *et al*. Polymorphisms of interferon-inducible genes OAS-1 and MxA associated with SARS in the Vietnamese population. *Biochem. Biophys. Res. Commun*. 329, 1234–1239 (2005).

45. Cheng, G. *et al.* Pharmacologic activation of the innate immune system to prevent respiratory viral infections. *Am. J. Respir. Cell Mol. Biol*. 45, 480–488 (2011).

46. Harari, D., Orr, I., Rotkopf, R., Baranzini, S. E. & Schreiber, G. A robust type I interferon gene signature from blood RNA defines quantitative but not qualitative differences

between three major IFNβ drugs in the treatment of multiple sclerosis. *Hum. Mol. Genet*. 24, 3192–3205 (2014).

47. Lin, F. & Young, H. A. Interferons: success in anti-viral immunotherapy. *Cytokine Growth Factor Rev*. 25, 369–376 (2014).

48. Arabi, Y. M. *et al*. Interferon beta-1b and lopinavir–ritonavir for Middle East respiratory syndrome. *N. Engl. J. Med*. 383, 1645–1656 (2020).

49. WHO Solidarity Trial Consortium. Repurposed antiviral drugs for COVID-19—interim WHO SOLIDARITY trial results. *N. Engl. J. Med*. 384, 497–511 (2021).

50. Monk, P. D. *et al*. Safety and efficacy of inhaled nebulised interferon beta-1a (SNG001) for treatment of SARS-CoV-2 infection: a randomised, double-blind, placebo-controlled, phase 2 trial. *Lancet Respir. Med*. 9, 196–206 (2020).

51. Wood, E. R. *et al*. The role of phosphodiesterase 12 (PDE12) as a negative regulator of the innate immune response and the discovery of antiviral inhibitors. *J. Biol. Chem*. 290, 19681–19696 (2015).

52. Silverman, R. H. & Weiss, S. R. Viral phosphodiesterases that antagonize double-stranded RNA signaling to RNase L by degrading 2-5A. *J. Interferon Cytokine Res*. 34, 455–463 (2014).

53. Zhao, L. *et al*. Antagonism of the interferon-induced OAS-RNase L pathway by murine coronavirus ns2 protein is required for virus replication and liver pathology. *Cell Host Microbe* 11, 607–616 (2012).

54. Zhang, R. *et al*. Homologous 2′,5′-phosphodiesterases from disparate RNA viruses antagonize antiviral innate immunity. *Proc. Natl Acad. Sci. USA* 110, 13114–13119 (2013).

55. Li, Y. *et al*. SARS-CoV-2 induces double-stranded RNA-mediated innate immune responses in respiratory epithelial derived cells and cardiomyocytes. Preprint at *bioRxiv* https://doi.org/10.1101/2020.09.24.312553 (2020).

56. Burgess, S., Davies, N. M. & Thompson, S. G. Bias due to participant overlap in two-sample Mendelian randomization. *Genet. Epidemiol*. 40, 597–608 (2016).

57. Chang, C. C. *et al*. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4, 7 (2015).

58. Auton, A. *et al*. A global reference for human genetic variation. *Nature* 526, 68–74 (2015).

59. Hemani, G. *et al*. The MR-Base platform supports systematic causal inference across the human phenome. *Elife* 7, e34408 (2018).

60. Candia, J. *et al*. Assessment of variability in the SOMAscan assay. *Sci. Rep*. 7, 14248 (2017).

61. SomaLogic. Short Technical Note. https://somalogic.com/wp-content/uploads/2019/07/Short-Technical-Note-SOMAmer-specificity.pdf (2019).

62. Centers for Disease Control and Prevention. Duration of isolation and precautions for adults with COVID-19. https://www.cdc.gov/coronavirus/2019-ncov/hcp/duration-isolation.html (2021).

63. Cevik, M. *et al.* SARS-CoV-2, SARS-CoV, and MERS-CoV viral load dynamics, duration of viral shedding, and infectiousness: a systematic review and meta-analysis. *Lancet Microbe* 2, e13–e22 (2020).

## 5.10 Supplemental data

Supplementary Tables and Figures can be downloaded from the open access publication Zhou, Bulter-Laporte, Nakanishi *et al*. in *Nat Med* available here:

https://www.nature.com/articles/s41591-021-01281-1#Sec20

## Chapter 6: General Discussion

The main purpose of this thesis was to advance the knowledge of the clinical implications of genetic determinants of respiratory diseases. In this doctoral thesis, I used sophisticated methods of genetic epidemiology, bioinformatics, and statistical genetics. This thesis represents a leap forward in clinical applications of genetic information in the field of multiple respiratory diseases, including AATD, IPF, and COVID-19. In Chapters 2 and 3, I explored the clinical impact of the genetic determinants for AATD and COVID-19 severity, using large-scale cohorts. In Chapters 4 and 5, we used the recent genome-wide association studies (GWASs) for proteome as exposure and those for respiratory diseases as outcomes in a Mendelian Randomization (MR) design. We identified potential causal circulating proteins that influence the IPF susceptibility, and the severity and susceptibility of COVID-19. Below, we discuss the strengths or shortcomings of each chapter.

In Chapter 2, I explored the clinical impact of the AATD-genotype (PI*ZZ) in the *SERPINA1* gene. We found that the vast majority of individuals with PI*ZZ were not diagnosed as having AATD in UKB. Yet, these individuals had substantially increased odds of respiratory symptoms, diseases, and all-cause mortality. We estimated that ~17000 individuals in the UK carry the PI*ZZ genotype. Thus, while the proportion of all British individuals who could be detected through genotyping efforts is small, the absolute number is not. Our study highlights the potential utility of population screening of this genotype, given its striking effects and the absolute number of individuals, who may otherwise suffer from disease burden without proper diagnosis and treatments. The major strength of the current study is that it is one of the largest studies to assess the effects of the *SERPINA1* genotype status on multiple health conditions in a single large population cohort. A prior family-based study included non-index family members with undiagnosed PI*ZZ individuals(39), which could reflect the effects of

other shared genetic factors. Since the participants in UK Biobank were recruited without regard to their symptoms or diseases, our study is less prone to such biases. The main limitation of this study is that UK Biobank is not representative of the general population as there is well-documented evidence of a "healthy volunteer' bias(40). Another shortcoming is that our study is based on the disease ascertainment using electrical health records, which could perhaps significantly underestimate the prevalence. There are no AAT measurements available in UK Biobank, so we could not test whether people with high-risk genotypes had low levels of plasma AAT.

In Chapter 3, we evaluated the major common genetic risk for severe COVID-19 on chromosome 3, which was tagged by rs10490770 C allele. Combining individual-level clinical and genomic data from 13,888 individuals ascertained for COVID-19 outcomes from 17 cohorts in 9 countries, we found that the risk allele was strongly associated with COVID-19–related mortality and clinical complications, such as respiratory failure and VTE. We also found that risk allele carriers aged ≤60 years had higher odds of death or severe respiratory failure (odds ratio [OR]: 2.7) compared with those >60 years (OR: 1.5). This risk variant improved the prediction of severe disease similarly to most clinical risk factors. The risk allele is common. We estimated that 14.4% of individuals of European ancestry are risk allele carriers at rs10490770. Further, 9.5% of admixed Americans, 2.4% of Africans, 47.1% of South Asians, and 0.4% of East Asians are risk allele carriers. Consequently, a large proportion of humans carry this risk factor. The major strength of this study is the large-scale aggregation of individual-level clinical and genotype data from multiple cohorts from diverse countries. Due to the nature of the heterogeneity of health care systems, our data from multiple countries substantially increases the generalizability of our research findings. Nevertheless, the dynamic variability in COVID-19 death rate, due to the different waves of strains and the different vaccination coverage rates(7), has made it particularly challenging to

generalize all results in the early pandemic to the current post-pandemic era in the COVID-19 research. Thus, future studies should re-evaluate the role of this major genetic risk or polygenic risk scores in COVID-19 severity prediction in the present situation of infection.

Importantly, the index genotypes both in Chapters 2 and 3 were not very rare (PI*Z allele: 0.03% and rs10490770: 15%, respectively). This implies that the absolute number of individuals who may benefit from the genotype information is substantial. To implement such genomics-guided clinical management in real-world settings, we further need to test the cost-effectiveness of the population-level screening in clinical trials. Moreover, as these are genetic studies with potential clinical implications, future efforts are needed to address the issue of incidental findings, such as applying the American College of Medical Genetics and Genomics(41) recommendations as to how to report secondary findings and setting up a proper genetic counseling system.

In Chapters 4 and 5, we applied MR to the large-scale pQTL GWASs, and the largest GWASs on IPF and the COVID-19 outcomes, to identify potential causal plasma proteins for disease susceptibility and severity, by efficiently scanning hundreds of proteins. MR is a well-established study design, which typically overcomes the bias from confounding and reverse causation of the observational studies. In Chapter 4, MR analyses of 834 proteins found that a 1 SD increase in circulating was associated with a reduced risk of IPF. FUT3 signals colocalised with IPF signals, with posterior probabilities of a shared genetic signal of 99.9%, and further transcriptomic investigations supported the protective effects of *FUT3* for IPF. In Chapter 5, in up to 14,134 cases and 1.2 million controls, we found that an SD increase in OAS1 levels was associated with reduced COVID-19 death or ventilation, hospitalization, and susceptibility. Measuring OAS1 levels in 504 individuals, we found that

higher plasma OAS1 levels in a non-infectious state were associated with reduced COVID-19 susceptibility and severity.

Given that the cost of measuring hundreds of proteins in adequately powered cohort studies involving samples collected years before disease onset is currently prohibitive, our approach provides an opportunity to prioritise candidate causal protein biomarkers by repurposing available data from large GWASs, which is the major strength of our analyses. It is important to underline that the potential causal proteins we identified (FUT3 and OAS1) were not exclusively responsible for the diseases, given the multi-factorial nature of IPF and COVID-19. For example, there is well-known evidence that telomere-related genes and surfactant proteins have important roles in IPF pathogenesis(16, 42). Similarly, members of the Toll-like receptor group such as *TLR7* and type 1 interferons are known to be the key players in host defense from SARS-CoV-2 infection. However, we did not identify such proteins in our pipeline, as we focused on circulating proteins; extracellular or secreted forms of proteins in the blood. Since respiratory diseases occur particularly in the lungs, it is valuable to understand the causal molecules in the lungs. Nevertheless, we focused only on circulating proteins, as we thought circulating proteins are an attractive source of biomarkers and therapeutic targets since they are easy to measure from blood, are more stable than mRNA, but are still able to target specific genes. MR studies for circulating biomarkers have often replicated or predicted the results of large-scale randomised controlled trials of pharmacological interventions to change biomarker levels(43–45). Thus, our MR analyses have direct translational relevance.

Our MR analyses for IPF and COVID-19 also expanded the knowledge of the disease pathophysiology. Although it is still unclear how FUT3 may influence IPF risk, the fucosyltransferases encoded by *FUT3* catalyse the formation of α-(1,4)-fucosylated

glycoconjugates and allow post-translational modification that attaches fucose residues to polysaccharides, called fucosylation(46). Fucosylation partly determines the heterogeneity of mucin size and charge, which are highly expressed in epithelial cells in the lungs. OAS1 is a double-stranded RNA (dsRNA) sensor capable of activating ribonuclease L (RNase L)(47). A recent functional study also demonstrated that prenylation of OAS1 appears to be necessary for dsRNA sensing of SARS-CoV-2(48). Collectively, our MR analyses supported targets with direct functional relevance to the diseases.

MR has important methodological limitations. MR rests on several assumptions(49), the most problematic being horizontal pleiotropy of the genetic instruments (wherein the genotype influences the outcome, independently of the exposure). Although we tried to avoid this bias by using genetic variants that influence circulating protein levels that are adjacent to the gene that encodes the circulating protein through the use of *cis*- pQTLs, we could not eradicate the possibility of this bias.

Lastly, the major limitation of all of the works described in Chapters 2 to 5 is that the genetic analyses were predominantly performed on individuals of European descent. Such an imbalanced abundance of European-descent studies may lead to the poor generalizability of genetic studies across populations(50). Thus, there is an urgent need to capture ancestral diversity in genetic studies to mitigate the potential health disparity in clinical translation of genetic findings. To expand our knowledge of the genetic basis of ILDs by capturing the diversity in populations, we aimed to perform multi-ethnic GWASs for ILDs. This is an ongoing project, as described in Appendix 3.

## Chapter 7: Conclusions and Future Directions

This thesis was an exploration of how to translate the genetic determinants of respiratory diseases into the clinical context. The obtained findings demonstrate the clinical utility of genetic information and the causal circulating proteins for diseases occurring in the lungs. Several future aims can be suggested to continue this work.

In Chapters 2 and 3, we addressed the potential clinical implications of genetic information. The important step toward genomics-guided clinical management is to test the cost-effectiveness of the population-level genetic screening. In addition, the incidental findings should be properly communicated to people by trained genetic counselors, applying the standards of American College of Medical Genetics and Genomics recommendations(51).

In Chapters 4 and 5, we performed MR analyses to identify causal circulating proteins for respiratory diseases. Although MR is a powerful tool to efficiently scan hundreds of proteins that could identify reliable therapeutic targets with a causal relationship with diseases, MR could only serve as a hypothesis-generating tool. Thus, further functional validation to understand the pathophysiology of how these proteins cause the diseases, and to test whether the inhibition of these proteins is effective as a treatment strategy.

Chapters 2 to 5 can be continued with trans-ancestry analyses. We restricted our analyses to the participants of European descent in all the analyses, to reduce the risk of confounding due to population stratification(52). In Chapter 2 and Chapter3, we tried to expand the analyses to non-European ancestries, however, these efforts have lacked statistical power to draw meaningful conclusions. To expand our knowledge of the genetic basis of ILDs by capturing the diversity in populations, we aimed to perform multi-ethnic GWASs for ILDs. This is

described in the proposal featured in Appendix 3. While we are to perform a trans-ancestry meta-analysis of GWAS for ILDs, the data used are still predominated by European ancestry. Further effort should be taken to build a well-powered biobank in non-European ancestries, such as BBJ(53), the Chinese Kadoorie Biobank(54), and Million Veteran Program(55).

Functional experiments are also the key step toward a deeper understanding of the biology of diseases. While such studies are on their way, the following research should be investigated to understand gene function in disease models: 1) Generation of more functional genomics data in various cell types in lungs (e.g. alveolar epithelial cells, ciliated cells, and basal cells); 2) Generation of high-throughput pooled genome-editing. Research on respiratory diseases relies on the isolation of primary cells from explanted lungs or the use of immortalized cells, which are both limited in their capacity to represent the genomic and phenotypic variability among the population. The use of patient-specific induced pluripotent cells (iPSC) is another emerging path to generating disease models that could represent the disease variability(56).

Lastly, this doctoral thesis mainly focused on the relationships between DNA (genome) and humans (phenome), with a slight exploration of transcriptome and proteome. While strong associations have been found between genome and phenome, there is still a large gap in how the genome influences the phenome. By incorporating other high dimensional omics data, such as single-cell RNA sequencing and spatial transcriptomic profiling(57), we could understand how genes influence transcripts, proteins, metabolites, and ultimately, phenotypes in a single-cell/cell-specific resolution.

In summary, this doctoral thesis provided a novel contribution to the field of genetics in respiratory medicine, by demonstrating potential opportunity to realize clinical benefits of

emerging worldwide genomic efforts and by identifying potentially druggable disease-influencing plasma proteins.

**Chapter 8 : Master reference list**


1. Sullivan J, Pravosud V, Mannino DM, Siegel K, Choate R, Sullivan T. National and state estimates of COPD morbidity and mortality - United States, 2014-2015. *Chronic Obstr Pulm Dis* 2018;5:324–333.

2. Vogelmeier CF, Criner GJ, Martinez FJ, Anzueto A, Barnes PJ, Bourbeau J, Celli BR, Chen R, Decramer M, Fabbri LM, Frith P, Halpin DMG, Varela MVL, Nishimura M, Roche N, Rodriguez-Roisin R, Sin DD, Singh D, Stockley R, Vestbo J, Wedzicha JA, Agustí A. Global strategy for the diagnosis, management, and prevention of chronic obstructive lung disease 2017 report. *Am J Respir Crit Care Med* 2017;195:557–582.

3. World Health Organization. Global surveilance, preventiona and control of CHRONIC RESPIRATORY DISEASES A comprehensive approach. *WHO Libr* 2007;1–37.doi:ISBN 978 92 4 156346 8.

4. Richeldi L, Collard HR, Jones MG. Idiopathic pulmonary fibrosis. *Lancet* 2017;389:1941–1952.

5. Hutchinson J, Fogarty A, Hubbard R, McKeever T. Global incidence and mortality of idiopathic pulmonary fibrosis: a systematic review. *Eur Respir J* 2015;46:795–806.

6. Vancheri C, Failla M, Crimi N, Raghu G. Idiopathic pulmonary fibrosis: A disease with similarities and links to cancer biology. *Eur Respir J* 2010;35:496–504.

7. WHO Coronavirus (COVID-19) Dashboard | WHO Coronavirus (COVID-19) Dashboard With Vaccination Data. at <https://covid19.who.int/>.

8. Buitrago-Garcia D, Egli-Gany D, Counotte MJ, Hossmann S, Imeri H, Ipekci AM, Salanti G, Low N. Occurrence and transmission potential of asymptomatic and presymptomatic SARSCoV-2 infections: A living systematic review and meta-analysis. *PLoS Med* 2020;

9. Lamprecht B, McBurnie MA, Vollmer WM, Gudmundsson G, Welte T, Nizankowska-Mogilnicka E, Studnicka M, Bateman E, Anto JM, Burney P, Mannino DM, Buist SA. COPD in never smokers: Results from the population-based burden of obstructive lung disease study. *Chest* 2011;139:752–763.

10. Silverman EK, Pierce JA, Province MA, Rao ; D C, Campbell EJ. Variability of Pulmonary Function in Alpha-1-Antitrypsin Deficiency: Clinical Correlates.

11. Silverman EK, Sandhaus RA. Alpha1-Antitrypsin Deficiency. *http://dx.doi.org/101056/NEJMcp0900449* 2009;360:2749–57.

12. Stoller JK, Aboussouan LS. A Review of α1-Antitrypsin Deficiency. *https://doi.org/101164/rccm201108-1428CI* 2012;185:246–259.

13. Dahl M, Hersh CP, Ly NP, Berkey CS, Silverman EK, Nordestgaard BG. The protease inhibitor PI*S allele and COPD: a meta-analysis. *Eur Respir J* 2005;26:67–76.

14. Nathan N, Giraud V, Picard C, Nunes H, Moal FD Le, Copin B, Galeron L, De Ligniville A, Kuziner N, Reynaud-Gaubert M, Valeyre D, Couderc LJ, Chinet T, Borie R, Crestani B, Simansour M, Nau V, Tissier S, Duquesnoy P, Mansour-Hendili L, Legendre M, Kannengiesser C, Coulomb-L'Hermine A, Gouya L, Amselem S, Clement A. Germline SFTPA1 mutation in familial idiopathic interstitial pneumonia and lung cancer. *Hum Mol Genet* 2016;25:1457–1467.

15. Wang Y, Kuan PJ, Xing C, Cronkhite JT, Torres F, Rosenblatt RL, DiMaio JM, Kinch LN, Grishin N V., Garcia CK. Genetic Defects in Surfactant Protein A2 Are Associated with Pulmonary Fibrosis and Lung Cancer. *Am J Hum Genet* 2009;84:52–59.

16. Stuart BD, Choi J, Zaidi S, Xing C, Holohan B, Chen R, Choi M, Dharwadkar P, Torres F, Girod CE, Weissler J, Fitzgerald J, Kershaw C, Klesney-Tait J, Mageto Y,

Shay JW, Ji W, Bilguvar K, Mane S, Lifton RP, Garcia CK. Exome sequencing links mutations in PARN and RTEL1 with familial pulmonary fibrosis and telomere shortening. *Nat Genet 2015 475* 2015;47:512–517.

17. Allen RJ, Guillen-Guio B, Oldham JM, Ma SF, Dressen A, Paynton ML, Kraven LM, Obeidat M, Li X, Ng M, Braybrooke R, Molina-Molina M, Hobbs BD, Putman RK, Sakornsakolpat P, Booth HL, Fahy WA, Hart SP, Hill MR, Hirani N, Hubbard RB, McAnulty RJ, Millar AB, Navaratnam V, Oballa E, Parfrey H, Saini G, Whyte MKB, Zhang Y, *et al.* Genome-wide association study of susceptibility to idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2020;201:564–574.

18. Fingerlin TE, Murphy E, Zhang W, Peljto AL, Brown KK, Steele MP, Loyd JE, Cosgrove GP, Lynch D, Groshong S, Collard HR, Wolters PJ, Bradford WZ, Kossen K, Seiwert SD, du Bois RM, Garcia CK, Devine MS, Gudmundsson G, Isaksson HJ, Kaminski N, Zhang Y, Gibson KF, Lancaster LH, Cogan JD, Mason WR, Maher TM, Molyneaux PL, Wells AU, *et al.* Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis. *Nat Genet 2013 456* 2013;45:613–620.

19. Docherty AB, Harrison EM, Green CA, Hardwick HE, Pius R, Norman L, Holden KA, Read JM, Dondelinger F, Carson G, Merson L, Lee J, Plotkin D, Sigfrid L, Halpin S, Jackson C, Gamble C, Horby PW, Nguyen-Van-Tam JS, Ho A, Russell CD, Dunning J, Openshaw PJM, Baillie JK, Semple MG. Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study. *BMJ* 2020;369:.

20. Van Der Made CI, Simons A, Schuurs-Hoeijmakers J, Van Den Heuvel G, Mantere T, Kersten S, Van Deuren RC, 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, De Mast Q, Bleeker-Rovers CP,

Joosten LAB, Yntema HG, Gilissen C, Nelen M, Van Der Meer JWM, Brunner HG, Netea MG, Van De Veerdonk FL, *et al.* Presence of Genetic Variants among Young Men with Severe COVID-19. *JAMA - J Am Med Assoc* 2020;324:663–673.

21. The Severe Covid-19 GWAS Group. Genomewide Association Study of Severe Covid-19 with Respiratory Failure. *N Engl J Med* 2020;NEJMoa2020283.doi:10.1056/NEJMoa2020283.

22. Pairo-Castineira E, Clohisey S, Klaric L, Bretherick AD, Rawlik K, Pasko D, Walker S, Parkinson N, Fourman MH, Russell CD, Furniss J, Richmond A, Gountouna E, Wrobel N, Harrison D, Wang B, Wu Y, Meynert A, Griffiths F, Oosthuyzen W, Kousathanas A, Moutsianas L, Yang Z, Zhai R, Zheng C, Grimes G, Beale R, Millar J, Shih B, *et al.* Genetic mechanisms of critical illness in Covid-19. *Nature* 2020;1–1.doi:10.1038/s41586-020-03065-y.

23. COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19. *Nature* 2021;1–8.doi:10.1038/s41586-021-03767-x.

24. Niemi MEK, Daly MJ, Ganna A. The human genetic epidemiology of COVID-19. *Nat Rev Genet 2022* 2022;1–14.doi:10.1038/s41576-022-00478-5.

25. Downes DJ, Cross AR, Hua P, Roberts N, Schwessinger R, Cutler AJ, Munis AM, Brown J, Mielczarek O, de Andrea CE, Melero I, Gill DR, Hyde SC, Knight JC, Todd JA, Sansom SN, Issa F, Davies JOJ, Hughes JR. Identification of LZTFL1 as a candidate effector gene at a COVID-19 risk locus. *Nat Genet 2021 5311* 2021;53:1606–1615.

26. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, Hirschhorn JN. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet 2008 95* 2008;9:356–369.

27. Rose NC, Kaimal AJ, Dugoff L, Norton ME. Screening for Fetal Chromosomal Abnormalities: ACOG Practice Bulletin, Number 226. *Obstet Gynecol* 2020;136:e48–e69.

28. Hu C, Hart SN, Polley EC, Gnanaolivu R, Shimelis H, Lee KY, Lilyquist J, Na J, Moore R, Antwi SO, Bamlet WR, Chaffee KG, DiCarlo J, Wu Z, Samara R, Kasi PM, McWilliams RR, Petersen GM, Couch FJ. Association Between Inherited Germline Mutations in Cancer Predisposition Genes and Risk of Pancreatic Cancer. *JAMA* 2018;319:2401–2409.

29. Ackerman MJ, Priori SG, Willems S, Berul C, Brugada R, Calkins H, Camm AJ, Ellinor PT, Gollob M, Hamilton R, Hershberger RE, Judge DP, Le Marec H, McKenna WJ, Schulze-Bahr E, Semsarian C, Towbin JA, Watkins H, Wilde A, Wolpert C, Zipes DP. HRS/EHRA Expert Consensus Statement on the State of Genetic Testing for the Channelopathies and Cardiomyopathies: This document was developed as a partnership between the Heart Rhythm Society (HRS) and the European Heart Rhythm Association (EHRA). *Hear Rhythm* 2011;8:1308–1339.

30. Nelson MR, Tipney H, Painter JL, Shen J, Nicoletti P, Shen Y, Floratos A, Sham PC, Li MJ, Wang J, Cardon LR, Whittaker JC, Sanseau P. The support of human genetic evidence for approved drug indications. *Nat Genet* 2015;47:856–860.

31. Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K, Kochi Y, Ohmura K, Suzuki A, Yoshida S, Graham RR, Manoharan A, Ortmann W, Bhangale T, Denny JC, Carroll RJ, Eyler AE, Greenberg JD, Kremer JM, Pappas DA, Jiang L, Yin J, Ye L, Su DF, Yang J, Xie G, Keystone E, Westra HJ, Esko T, *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nat 2013 5067488* 2013;506:376–381.

32. Shrine N, Guyatt AL, Erzurumluoglu AM, Jackson VE, Hobbs BD, Melbourne CA, Batini C, Fawcett KA, Song K, Sakornsakolpat P, Li X, Boxall R, Reeve NF, Obeidat

M, Zhao JH, Wielscher M, Weiss S, Kentistou KA, Cook JP, Sun BB, Zhou J, Hui J, Karrasch S, Imboden M, Harris SE, Marten J, Enroth S, Kerr SM, Surakka I, *et al.* New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. *Nat Genet* 2019;51:481–493.

33. Barrecheguren M, Monteagudo M, Simonet P, Llor C, Rodriguez E, Ferrer J, Esquinas C, Miravitlles M. Diagnosis of alpha-1 antitrypsin deficiency: a population-based study. *Int J Chron Obstruct Pulmon Dis* 2016;11:999.

34. Soriano JB, Lucas SJ, Jones R, Miravitlles M, Carter V, Small I, Price D, Mahadeva R. Trends of testing for and diagnosis of α1-antitrypsin deficiency in the UK: more testing is needed. *Eur Respir J* 2018;52:.

35. Aboussouan LS, Stoller JK. Detection of alpha-1 antitrypsin deficiency: A review. *Respir Med* 2009;103:335–341.

36. Bovijn J, Lindgren CM, Holmes M V. Genetic variants mimicking therapeutic inhibition of IL-6 receptor signaling and risk of COVID-19. *Lancet Rheumatol* 2020;2:e658–e659.

37. Butler-Laporte G, Nakanishi T, Mooser V, Renieri A, Amitrano S, Zhou S, Chen Y, Forgetta V, Richards JB. The effect of angiotensin-converting enzyme levels on COVID-19 susceptibility and severity: a Mendelian randomization study. *Int J Epidemiol* 2021;50:75–86.

38. Butler-Laporte G, Nakanishi T, Mooser V, Morrison DR, Abdullah T, Adeleye O, Mamlouk N, Kimchi N, Afrasiabi Z, Rezk N, Giliberti A, Renieri A, Chen Y, Zhou S, Forgetta V, Richards JB. Vitamin D and COVID-19 susceptibility and severity in the COVID-19 Host Genetics Initiative: A Mendelian randomization study. *PLOS Med* 2021;18:e1003605.

39. DeMeo DL, Sandhaus RA, Barker AF, Brantly ML, Eden E, McElvaney NG, Rennard S, Burchard E, Stocks JM, Stoller JK, Strange C, Turino GM, Campbell EJ, Silverman EK. Determinants of airflow obstruction in severe alpha-1-antitrypsin deficiency. *Thorax* 2007;62:806–813.

40. Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, Collins R, Allen NE. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol* 2017;186:1026–1034.

41. Miller DT, Lee K, Chung WK, Gordon AS, Herman GE, Klein TE, Stewart DR, Amendola LM, Adelman K, Bale SJ, Gollob MH, Harrison SM, Hershberger RE, McKelvey K, Richards CS, Vlangos CN, Watson MS, Martin CL. ACMG SF v3.0 list for reporting of secondary findings in clinical exome and genome sequencing: a policy statement of the American College of Medical Genetics and Genomics (ACMG). *Genet Med 2021 238* 2021;23:1381–1390.

42. Adegunsoye A, Vij R, Noth I. Integrating Genomics Into Management of Fibrotic Interstitial Lung Disease. *Chest* 2019;

43. Manousaki D, Mokry LE, Ross S, Goltzman D, Brent Richards J. Mendelian Randomization Studies Do Not Support a Role for Vitamin D in Coronary Artery Disease. *Circ Cardiovasc Genet* 2016;9:349–356.

44. Manson JAE, Cook NR, Lee IM, Christen W, Bassuk SS, Mora S, Gibson H, Gordon D, Copeland T, D'Agostino D, Friedenberg G, Ridge C, Bubes V, Giovannucci EL, Willett WC, Buring JE. Vitamin D supplements and prevention of cancer and cardiovascular disease. *N Engl J Med* 2019;380:33–44.

45. Holmes M V., Smith GD. Revealing the effect of CETP inhibition in cardiovascular disease. *Nat Rev Cardiol 2017 1411* 2017;14:635–636.

46. Dupuy F, Germot A, Marenda M, Oriol R, Blancher A, Julien R, Maftah A. α1,4-fucosyltransferase activity: A significant function in the primate lineage has appeared twice independently. *Mol Biol Evol* 2002;19:815–824.

47. Han Y, Donovan J, Rath S, Whitney G, Chitrakar A, Korennykh A. Structure of human RNase L reveals the basis for regulated RNA decay in the IFN response. *Science (80- )* 2014;343:1244–1248.

48. Wickenhagen A, Sugrue E, Lytras S, Kuchi S, Noerenberg M, Turnbull ML, Loney C, Herder V, Allan J, Jarmson I, Cameron-Ruiz N, Varjak M, Pinto RM, Lee JY, Iselin L, Palmalux N, Stewart DG, Swingler S, Greenwood EJD, Crozier TWM, Gu Q, Davies EL, Clohisey S, Wang B, Costa FTM, Santana MF, de Lima Ferreira LC, Murphy L, Fawkes A, *et al.* A prenylated dsRNA sensor protects against severe COVID-19. *Science (80- )* 2021;374:.

49. Davey Smith G, Davies NM, Dimou N, Egger M, Gallo V, Golub R, Higgins JP, Langenberg C, Loder EW, Brent Richards J, Richmond RC, Skrivankova VW, Swanson SA, Timpson NJ, Tybjaerg-Hansen A, VanderWeele TJ, Woolf BA, Yarmolinsky J. STROBE-MR: Guidelines for strengthening the reporting of Mendelian randomization studies. 2019;doi:10.7287/peerj.preprints.27857v1.

50. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet* 2019;51:584–591.

51. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL. Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 2015;17:405.

52. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. *Nat Rev Genet* 2010;11:459.

53. Nagai A, Hirata M, Kamatani Y, Muto K, Matsuda K, Kiyohara Y, Ninomiya T, Tamakoshi A, Yamagata Z, Mushiroda T, Murakami Y, Yuji K, Furukawa Y, Zembutsu H, Tanaka T, Ohnishi Y, Nakamura Y, Kubo M, Shiono M, Misumi K, Kaieda R, Harada H, Minami S, Emi M, Emoto N, Daida H, Miyauchi K, Murakami A, Asai S, *et al.* Overview of the BioBank Japan Project: Study design and profile. *J Epidemiol* 2017;27:S2–S8.

54. Chen Z, Chen J, Collins R, Guo Y, Peto R, Wu F, Li L, Lancaster G, Yang X, Williams A, Smith M, Yang L, Chang Y, Guo Y, Zhao G, Bian Z, Wu L, Hou C, Pang Z, Wang S, Zhang Y, Zhang K, Liu S, Zhao Z, Liu S, Pang Z, Feng W, Wu S, Yang L, *et al.* China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *Int J Epidemiol* 2011;40:1652.

55. Gaziano JM, Concato J, Brophy M, Fiore L, Pyarajan S, Breeling J, Whitbourne S, Deen J, Shannon C, Humphries D, Guarino P, Aslan M, Anderson D, Lafleur R, Hammond T, Schaa K, Moser J, Huang G, Muralidhar S, Przygodzki R, O'Leary TJ. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol* 2016;70:214–223.

56. Calvert BA, Ryan AL. Application of iPSC to Modelling of Respiratory Diseases. *Adv Exp Med Biol* 2019;1237:1–16.

57. Merritt CR, Ong GT, Church SE, Barker K, Danaher P, Geiss G, Hoang M, Jung J, Liang Y, McKay-Fleisch J, Nguyen K, Norgaard Z, Sorg K, Sprague I, Warren C, Warren S, Webster PJ, Zhou Z, Zollinger DR, Dunaway DL, Mills GB, Beechem JM. Multiplex digital spatial profiling of proteins and RNA in fixed tissue. *Nat Biotechnol 2020 385* 2020;38:586–599.

58. Cottin V, Wollin L, Fischer A, Quaresma M, Stowasser S, Harari S. Fibrosing interstitial lung diseases: knowns and unknowns. *Eur Respir Rev* 2019;28:.

59. Steele MP, Speer MC, Loyd JE, Brown KK, Herron A, Slifer SH, Burch LH, Wahidi MM, John A. Phillips I, Sporn TA, McAdams HP, Schwarz MI, Schwartz DA. Clinical and Pathologic Features of Familial Interstitial Pneumonia. *https://doi.org/101164/rccm200408-1104OC* 2012;172:1146–1152.

60. Ley B, Newton CA, Arnould I, Elicker BM, Henry TS, Vittinghoff E, Golden JA, Jones KD, Batra K, Torrealba J, Garcia CK, Wolters PJ. The MUC5B promoter polymorphism and telomere length in patients with chronic hypersensitivity pneumonitis: an observational cohort-control study. *Lancet Respir Med* 2017;5:639–647.

61. Ley B, Torgerson DG, Oldham JM, Adegunsoye A, Liu S, Li J, Elicker BM, Henry TS, Golden JA, Jones KD, Dressen A, Yaspan BL, Arron JR, Noth I, Hoffmann TJ, Wolters PJ. Rare Protein-Altering Telomere-related Gene Variants in Patients with Chronic Hypersensitivity Pneumonitis. *https://doi.org/101164/rccm201902-0360OC* 2019;200:1154–1163.

# Appendices

## Appendix 1: Copyright Permissions

## Appendix 2: Ethics and related certificates

For Chapters 2 to 5, informed consent was obtained for each participant and was approved by each participating sites' regional ethical review board.

**Appendix 3: Trans-ancestry genome-wide association study to identify genetic determinants for respiratory diseases.**

**Title: Trans-ancestry genome-wide association study to identify genetic determinants for respiratory diseases.**

**Brief background and research goals**

Interstitial lung disease (ILD) is a heterogeneous assembly of diseases that specifically affect lung parenchyma and alveoli, characterized by worsening quality of life and decline in lung function(58). Idiopathic pulmonary fibrosis (IPF) is the most common and genetically explored type of ILDs with the median survival time from diagnosis being 3 to 5 years with few treatment options, which is worse than the prognosis of several types of cancers(6).

Owing to the prior family-based genetic studies for familial pulmonary fibrosis (FPF)(14–16) and large-scale genome-wide association studies (GWASs) for IPF(17, 18), both rare variants (e.g. in telomere-related genes and surfactant-associated protein genes) and common variants (e.g. in *MUC5B, DSP*, and telomere-related genes) have been discovered to predispose to IPF. However, genetic studies have been conducted predominantly in individuals of European descent, thus its generalizability to non-European ancestry is not guaranteed. Ultimately, such Eurocentric genetic study biases may exacerbate health disparities as clinical uses of genetic findings get widely implemented(50). Given that the incidence of IPF is lower in East Asia than in European countries(5), there could be some genetic diversity between ancestries.

Moreover, the genetic background of non-IPF ILDs is not well understood because of the predominance of IPF-oriented genetic studies. Some evidence points to a shared genetic basis of IPF and non-IPF ILDs. For instance, pathological heterogeneity was observed within

family members of FPF with numerous families having evidence of usual interstitial pneumonia (UIP)/IPF and non-specific interstitial pneumonia (NSIP) histopathology(59). Recent studies have demonstrated that those with non-IPF fibrosing ILDs, such as chronic hypersensitivity pneumonitis, also carried the IPF-associated *MUC5B* promoter variant(60) and rare variants in telomere-related genes(61).

To expand our knowledge, we try to

1. Screen putative pathogenic variants in candidate genes in the Japanese cohort (the Kyoto-ILD cohort; 52 FPF, 162 sporadic IPF [sIPF] cases with whole-genome sequencing [WGS] data).
2. Combine data of gene-based tests for ILD using individuals of Japanese and European descent in the Kyoto-ILD cohort and UKB, to identify rare genetic variants associated with disease risk across populations.
3. Perform genome-wide association meta-analysis of ILD, using the Kyoto-ILD cohort, UKB, Biobank Japan (BBJ)(53), FinnGen, and the Chicago/Colorado/UK study(17), to identify novel common genetic variants.
4. Establish polygenic risk scores (PRSs), a weighted sum of the effect sizes of common variants, for IPF to evaluate its associations with non-IPF-ILDs in the Kyoto-ILD cohort and UKB, where individual-level data were available.

**Appendix 4: Significant Contributions by the Author to Other Projects.**

**Peer-Reviewed Publications** * denotes equal contribution

**2022**

1. Huffman J E, Butler-Laporte G, Khan A, Pairo-Castineira E, Drivas TG, Peloso GM, **Nakanishi T**, Ganna A, Verma A, Baillie JK, Kiryluk K, Richards JB, Zeberg H, Multi-ancestry fine mapping implicates OAS1 splicing in risk of severe COVID-19, ***Nat. Genet***. 2022, 1–3 (2022).

**2021**

2. Butler-Laporte G*, **Nakanishi T***, Mooser V, Morrison DR, Abdullah T, Adeleye O, Mamlouk N, Kimchi N, Afrasiabi Z, Rezk N, Giliberti A, Renieri A, Chen Y, Zhou S, Forgetta V, Richards JB. Vitamin D and COVID-19 susceptibility and severity in the COVID-19 Host Genetics Initiative: A Mendelian randomization study. ***PLOS Med*** 2021;18:e1003605.
3. Butler-Laporte G*, **Nakanishi T***, Mooser V, Renieri A, Amitrano S, Zhou S, Chen Y, Forgetta V, Richards JB. The effect of angiotensin-converting enzyme levels on COVID-19 susceptibility and severity: a Mendelian randomization study. ***Int J Epidemiol*** 2021;50:75–86.
4. COVID-19 Host Genetics Initiative (**contributing author**). Mapping the human genetic architecture of COVID-19. ***Nature*** 2021;1–8.doi:10.1038/s41586-021-03767-x.
5. Povysil G, Butler-Laporte G, Shang N, Wang C, Khan A, Alaamery M, **Nakanishi T**, Zhou S, Forgetta V, Eveleigh RJM, Bourgey M, Aziz N, Jones SJM, Knoppers B, Scherer SW, Strug LJ, Lepage P, Ragoussis J, Bourque G, Alghamdi J, Aljawini N, Albes N, Al-Afghani HM, Alghamdi B, Almutairi MS, Mahmoud ES, Abu-Safieh L, El Bardisy H, Al Harthi FS, et al. Rare loss-of-function variants in type I IFN immunity genes are not associated with severe COVID-19. ***J Clin Invest*** 2021;131:

6. Brunet-Ratnasingham E, Anand SP, Gantner P, Dyachenko A, Moquin-Beaudry G, Brassard N, Beaudoin-Bussières G, Pagliuzza A, Gasser R, Benlarbi M, Point F, Prévost J, Laumaea A, Niessl J, Nayrac M, Sannier G, Orban C, Messier-Peet M, Butler-Laporte G, Morrison DR, Zhou S, **Nakanishi T**, Boutin M, Descôteaux-Dinelle J, Gendron-Lepage G, Goyette G, Bourassa C, Medjahed H, Laurent L, et al. Integrated immunovirological profiling validates plasma SARS-CoV-2 RNA as an early predictor of COVID-19 mortality. ***Sci Adv*** 2021;7:5629.
7. Kosmicki JA, Horowitz JE, Banerjee N, Lanche R, Marcketta A, Maxwell E, Bai X, Sun D, Backman JD, Sharma D, Kury FSP, Kang HM, O'Dushlaine C, Yadav A, Mansfield AJ, Li AH, Watanabe K, Gurski L, McCarthy SE, Locke AE, Khalid S, O'Keeffe S, Mbatchou J, Chazara O, Huang Y, Kvikstad E, O'Neill A, Nioi P, Parker MM, …, **Nakanishi T**, et al. Pan-ancestry exome-wide association analyses of COVID-19 outcomes in 586,157 individuals. ***Am J Hum Genet*** 2021;108:1350–1355.
8. Ikegami N, Nakajima N, Yoshizawa A, Handa T, Chen-Yoshikawa T, Kubo T, Tanizawa K, Ohsumi A, Yamada Y, Hamaji M, Nakajima D, Yutaka Y, Tanaka S, Watanabe K, Nakatsuka Y, Murase Y, **Nakanishi T**, Niwamoto T, Chin K, Date H, Hirai T. Clinical, radiological, and pathological features of idiopathic and secondary interstitial pneumonia cases with pleuroparenchymal fibroelastosis undergoing lung transplantation. ***Histopathology*** 2021;doi:10.1111/HIS.14595.

**2020**

9. The Severe Covid-19 GWAS Group (**contributing author**). Genomewide Association Study of Severe Covid-19 with Respiratory Failure. ***N Engl J Med*** 2020;NEJMoa2020283.doi:10.1056/NEJMoa2020283.
10. Butler-Laporte G, Kreuzer D, **Nakanishi T**, Harroud A, Forgetta V, Richards JB. Genetic Determinants of Antibody-Mediated Immune Responses to Infectious

Diseases Agents: A Genome-Wide and HLA Association Study. ***Open Forum Infect Dis*** 2020;7