An anonymized GWAS to urgently query host genetic predisposition to severe COVID-19 (SARS-CoV-2 infection) lung disease

The COVID-19 epidemic situation needs little introduction and represent a global world-wide emergency with mortality rates rapidly increasing in Europe and the US. Disease behavior is variable, with the majority of patients experiencing only mild symptoms or no symptoms at all. Some patients develop severe pulmonary affection, with aggressive and extensive inflammatory destruction of lung parenchyma and associated inflammatory responses and superinfections, driving large fractions of the COVID-19 related mortality. What exactly drives this development of severe lung disease remains unknown, but old age, obesity, diabetes and other co-morbidities increase the risk, while the role played by specific medications is still uncertain. Variation in virus genetics and patient immunology are also likely involved. As to the latter point, we hypothesize that host genetics may play a role in determining development of severe lung disease in SARS-CoV-2 infection.

Genome-wide association studies (GWAS) have been applied to decipher the genetic predisposition in thousands of disease traits since the study design was invented in 2005. The genetic signals detected vary from very strong effects that can be detected in a few hundred individuals, to very weak effects requiring cohorts of tens of thousands for detection. By 2020, the study design is now a robust, off-the-shelf, easy-to-perform industry-standard screening tool for genetic predisposition, even available through "consumer genetics" online-based companies. The study design is simple: testing for genetic variants throughout the genome (single nucleotide polymorphisms, SNPs) using SNP microarrays, comparing their frequencies in patients versus controls (or across other variables). For inflammatory phenotypes in particular, GWAS has proven an efficient tool, delineating hundreds of susceptibility loci in many conditions, some of which has provided novel and surprising disease insights.

GWAS serve two purposes. Most importantly they allow to determine biological factors involved in disease development, thus potentially guiding drug development and therapy. This would be particularly relevant during the current COVID-19 emergency, when hundreds of trials have begun and there is an urgent need to prioritize well-conducted collaborative studies based on robust pathophysiological data. Secondly, and increasingly popular, they allow for the calculation of a "polygenic risk score" to predict disease development. Both aspects appear crucial to clarify for COVID-19 lung disease: (a) are there genetic signatures suggesting which biological mechanisms are involved that may suggest relevant therapeutic approaches, and (b) can we predict those at risk (or those with very low risk)?

We hereby call for participation in an "open science", "superfast" and "supersimple" GWAS initiative to early on identify host genetic factors in severe COVID-19 lung disease, made possible through minimal effort anonymized sample collection ("extra blood tube day") combined with external (German/Norwegian) technical and financial support to perform all other work needed.

Hypothesis: Host genetic factors contribute to development of severe pulmonary disease in patients with SARS-CoV-2 infection.

Aims: The aims of the study are (i) to identify genetic factors that predispose or protect against development of severe pulmonary affection in SARS-CoV-2 infected patients, and (ii) to develop a polygenic risk score to identify individuals at particularly high or low risk for severe lung disease.

Purpose: The purpose of the study is to identify genetic factors that may educate the biological understanding of the development of severe lung disease in SARS-CoV-2 infection. Such an

understanding may inform treatment trials of potential utility towards more effective management of this patient group. A polygenic risk score may also be useful in identifying high/low risk patients, e.g. amongst health-care providers and vulnerable individuals with significant comorbidities.

Study design: The study is a case-control study. Cases will be defined by SARS-CoV-2 positive patient with severe lung affection defined by hospitalization and respiratory failure requiring support of any kind (ranging from O_2 support via non-invasive ventilation like CPAB/BiPAP/HFNC to full respirator/ECMO). Healthy control data are available from previous multi-ethnic assessments, and will be the source of the case-control comparison in this fast project. Later on, in due time, there will be more refined studies comparing individuals who have been infected by SARS-CoV-2 without developing clinically significant symptoms, patients with mild disease versus patients with severe disease, adjusting for multiple co-variables, but these studies will arrive too late to inform the peak of the epidemic. SAMPLE COLLECTION FOR THE PRESENT STUDY SHOULD THUS BE PERFORMED WITHIN 2-3 WEEKS.

Methods and statistical power considerations: One extra tube of EDTA blood is collected during routine biochemistry and frozen. Buffy-coat (or remains) are also suitable for DNA extraction. After appropriate courier shipment, DNA will be extracted using standard column-based methodology and SNP genotyping will be done using the Illumina Global Screening Array (GSA), for details see https://emea.illumina.com/products/by-type/microarray-kits/infinium-global-screening.html.
Statistical association analysis will be performed using logistic regression / linear mixed models based association methods with standard tools (e.g. SAIGE, BOLT-LMM, PLINK etc.)) and appropriate correction for population stratification confounders, and incorporating patient age and gender as co-variables. For strong genetic signals (e.g. HLA in many immune mediated diseases, IL28B in HCV infection, complement factor H polymorphism in age-related macular degeneration) only a few hundred patients have been required to achieve genome-wide significant findings. Our study is originally scoped toward 2000 cases, but if more samples are collected, the experiment will be expanded (for which there is both capacity and funding). Associations with very weak effect sizes may be missed with the crude approach, but for this study this compromise is acceptable to achieve feasibility and speed.

Sample identification: The study will by default be performed COMPLETELY ANONYMIZED (not deidentified/pseudonymized), with a minimum of information (age in years, type of respiratory support 1-4 – O2/CPAP-BIPAP-HFNC/ventilator/ECMO, gender), there will be applied NO personal identifiers and no clinical details that may lead to risk of post hoc deduction of ID. This complete anonymization, with no local list of re-connection to personal ID, is required to request exemption from informed consent, which is challenging in severely ill patients with advanced respiratory support (many will be sedated). Digit 1-2 [center no. – 01, 02, 03 etc.], digit 3-6 (patient no. – 0001, 0002, 0003 etc.), digit 7-8 (age – 42, 56, 75 etc.), digit 9 type of respiratory support (1 – O2 only, 2 – non-invasive ventilation, 3 ventilator, 4 – ECMO) digit 10 (gender – 1 female, 2 male) – leading to a 10 digit single identifier for each sample. Optional 11^{th} digit identifier (smoking, 1 no, 2 yes). Optional 12^{th} digit identifier (known pre-existing cardiovascular disease, 1 no, 2 yes). In respect of burdened clinicians, the 11^{th} and 12^{th} are optional, but would enhance confounder management in statistics.

Some centers have collected samples as part of internal / local research projects which involves other activities. For these centers, deidentified/pseudonymized samples will be allowed and handled, to allow for return of data and coupling to other variables in their research projects.

Execution: The execution of the project is built to allow clinicians working under stressful conditions as to be a “supersimple” and “superfast” undertaking. Upon establishing of necessary ethical

clearance and other necessary formalities (e.g. institutional material transfer agreements), a day is decided during which one extra tube of EDTA blood is drawn alongside routine biochemistry venipuncture **("the extra blood tube day")**. Tubes (ideally 5 ml but as little as 200 µl of EDTA blood are also sufficient) are anonymized and labeled with pre-printed stickers and frozen at -20^{o}C. Through relevant courier setup, blood is transferred to the Institute of Clinical and Molecular Biology (IKMB) in Kiel for DNA extraction and genotyping using the Illumina GSA. The lab holds accreditation for clinical grade analysis and serves also COVID-19 clinical testing for the relevant region (DIN EN ISO 15189, and has a 15 year leading reputation and experience in GWAS execution. Relevant post docs and bioinformaticians, accustomed to working with array genotype-data from multiple ethnicities will be liberated from existing duties to rapidly perform analysis, using genotype data from patients, and relevant control genotype data available through past and ongoing projects. We aim at collecting at least 2000 cases within the coming weeks, but upon achievement of a robust genetic signal (samples will be analyzed as batches come in), findings will be made publicly available and rapidly published in open access format with broad media channels communications. Lead clinical contributors at the forefront of the patient management will be guaranteed lead authorships to potential publications, or, alternatively if many centers and lead clinicians end up contributing, the possibility of a single banner authorship ("The COVID-19 rapid GWAS working group") with an alphabetical contributor list is the preferred alternative. Importantly, for all involved the academic merits associated with the work is irrelevant, the project is executed to rapidly achieve insights of relevance to managing patients and potentially saving lives.

Ethical considerations: The ethical implications to the proposal operates at several layers. Foremost, as the current epidemic is rapidly expanding, one could claim that an ethical obligation to as quickly as possible expand knowledge on why some patients develop severe lung disease during COVID-19 infection exists, so that this knowledge can benefit the rapidly increasing number of COVID-19 patients worldwide. Nevertheless, privacy and individual and family rights cannot be compromised, particularly in sensitive and stressful situations of severe disease with high mortality rates. A particular challenge in focusing on patients with the severe lung disease, is that they will be too sick to be competent for written informed consent, meaning a pathway for research that allows for exemption has to be delineated. Finally, as clinicians involved in patient management are under extreme pressure in the overcrowded hospital settings seen in many countries right now, any research effort has to impose minimal disturbance and disruption to clinical work. In the present proposal, we aim to account for all these concerns. First, to account for privacy, we will opt for a totally anonymized model with virtually no clinical data and total detachment from any patient identifier. Any data will still be handled in line with GDPR and highest standards in a laboratory used to handling clinical grade samples and data. This model, in our opinion, makes the exemption from written consent, which is anyways not practically feasible, acceptable. Second, we will opt for a model whereby implied clinicians will have minimal disruption by using the "one single day, one single EDTA blood tube extra during routine biochemistry"-concept, whilst the rest of the burden (and costs) for the assessment will be put on external actors (technical platform and analysts). Finally, inherent to the project is open sharing of the knowledge – as soon as a robust genetic signal has been detected, it will be publicly communicated and published in open access format. Individual level genotype data, however, cannot be made available to the public. Only summary statistics can be shared with third parties and leave the secure network of the Kiel diagnostics lab. At end of experiment, data will be returned to the participating institutions for potential further studies. Ethical clearances will be established per regional/national/institutional policies of each individual participant/contributor, and ethics committee recommendations and demands may vary accordingly. May ethics committees have fast-track for COVID-19 research projects.

Contact points: Project participation is open to anyone who wishes to contribute EDTA blood or DNA from anonymized patients with severe COVID-19 lung disease. Interested parties may contact national contact points, or the Norwegian/German coordinators as per preference. This mini-protocol is free for sharing, and anyone who wishes to do similar studies in the framework of other practical/financial settings are encouraged to do so – the important thing is to gather knowledge. Feel also free to evolve and improve and refine the protocol, with notification and a clear indication if this is done.

Tom Hemming Karlsen (contact point for practical information), Institute for Clinical Medicine and Department of Surgery, Inflammatory Diseases and Transplantation, Oslo University Hospital and University of Oslo, email: t.h.karlsen@medisin.uio.no, phone +47 91722353. Weblink to profile: https://www.ous-research.no/home/karlsen/Group%20members/6913.

Andre Franke (contact point for genotyping), Director of the Institute for Clinical and Molecular Biology, University of Kiel and University Clinic of Schleswig-Holstein, email address: a.franke@ikmb.uni-kiel.de, phone +49 1794851891. Weblink to profile, https://www.ikmb.uni-kiel.de/people/scientists/andre-franke

Luca Valenti, Department of Pathophysiology and Transplantation, University of Milan, and Translational Medicine – Biobank at the Department of Transfusion Medicine and Hematology, Fondazione IRCCS Ca' Granda Ospedale Maggiore IRCCS, Milano, Italy. Email: luca.valenti@unimi.it, phone +39 3381078229.

Jesús María Bañales, Department of Liver and Gastrointestinal Diseases, Donostia University Hospital, San Sebastian, Spain. Emalil: jesus.banales@biodonostia.org, phone: +34 627401179.

Pietro Invernizzi, San Gerardo Hospital, Monza, Italy. Email: pietro.invernizzi@unimib.it.

Javier Fernández, Hospital Clinic Barcelona, Spain. Email: JFDEZ@clinic.cat.

Agustín Albillos, Hospital Universitario Ramón y Cajal, Madrid, Spain. Email: agustin.albillos@uah.es.

Patrizio Burra, Padova University Hospital, Padova, Italy. Email: burra@unipd.it.

José Luis Del Pozo, University of Navarra Clinic, Pamplona, Spain. Email: jdelpozo@unav.es.

Manuel Romero Gomez, Hospital Virgen del Rocio, Seville, Spain. Email: mromerogomez@us.es.

Stefano Duga, Humanitas University, Milan, Italy. Email: stefano.duga@hunimed.eu.

Other countries and collaborators are more than free to join, write to any above to express your interest and we will keep updating this document as the confirmed contributors list evolves. We will also provide you with all necessary practical information, help with sample shipment details/equipment, and support formality clearances as needed.

Funding: The project is made possible through financial contributions through generic grants at coordinating institutions and a philanthropic donation from Stein Erik Hagen through Canica A/S.