Connecting European Cohorts to Increase Common and Effective Response to SARS- CoV-2 Pandemic

WPN°6 Deliverable N°6.2,

Dataset of human genome sequencing of patients with COVID-19 from retrospective cohorts

Beneficiary in charge

UNIBO, INSERM

Project Classification

Project Acronym:	ORCHESTRA
Project Title:	Connecting European Cohorts to Increase Common and Effective Response to SARS- CoV-2 Pandemic
Coordinator:	UNIVR
Grant Agreement Number:	101016167
Funding Scheme:	Horizon 2020
Start:	1st December 2020
Duration:	36 months
Website:	www.orchestra-cohort.eu
Email:	info@orchestra.eu

Document Classification

WP No:	WP6
Deliverable No:	D6.2
Title:	Genomic sequencing from retrospective cohorts completed
Lead Beneficiary:	UNIBO
Other Involved Beneficiaries:	INSERM
Nature:	Report
Dissemination Level:	Confidential
Due Delivery Date:	M9, postponed to M13
Submission Date:	27/01/2022
Justification of delay:	Sequencing out-sourcing providers were established through public tender procedures that lasted several months; approvals of human genomics in ethical protocols were delayed by several months.
Status:	In Progress
Version:	1.0
Author(s):	Laurent Abel, Tommaso Pippucci, Marco Seri, Valerio Carelli

ORCHESTRA has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101016167

History of Changes

Version	Date	Created/Modified by
0.1	23/12/2021	Laurent Abel, Tommaso Pippucci, Marco Seri, Valerio Carelli
0.2	27/01/2022	Laurent Abel, Tommaso Pippucci

ORCHESTRA has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101016167

Table of contents

Executive summary

This deliverable (Deliverable WP6.2) is part of the WP6 task on genomic sequencing of COVID-19 patients “with varying degrees of disease severity or asymptomatic” by UNIBO and INSERM. The purpose of the document is to report on the state of sequencing of retrospective cohorts so far. Here are described the cohorts on which this task has been performed and the number of samples that have been sent to sequencing. The target of this task is to reach the expected number of samples of genomic DNA of COVID-19 patients sent to either WES or WGS. The expected number of sequenced samples for the retrospective study is 360. Dissemination level is confidential.

Core content

Background

Recent research has shown that rare and common genetic variants are associated with severe COVID-19 disease; relevant pathways have been identified in interferon type-1 mediated immunity and others (Zhang 2020; Pairo-Castineira, 2021). In particular, rare variants in 8 genes involved in type-I IFN immunity were found to significantly contribute to individual predisposition to life-threatening COVID-19 pneumonia in about 3.5% of the studied patients (Zhang 2020). A further analysis showed that ~1% of male patients under the age of 60 years had X-linked recessive TLR7 deficiency (Asano 2021). Studying the host genome is therefore of major interest for identifying genetic factors of disease predisposition and the underlying mechanisms of pathogenicity.

Methods

Whole Exome Sequencing (WES) or Whole Genome Sequencing (WGS) is used to understand the role of coding (WES) and coding/non-coding variants (WGS). In the

ORCHESTRA has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101016167

retrospective study, sequencing of 360 patients “with varying degrees of disease severity or asymptomatic” is ongoing. High-quality variants are generated from the sequence reads obtained by WES/WGS using pipelines to obtain VCF files using the following criteria:

- Genotype-level parameters: minimum coverage, allelic balance, and genotype quality
- Variant-level parameters: GATK variant quality score recalibration (VQSR) and call rate
- Individual-level parameters: call rate, mean coverage, number of heterozygous calls, transition/transversion & snp/indel ratios
- Sex inference from the WES/WGS data and checking for consistency with reported sex
- Ethnic origin inference by means of reported ethnicity and principal component analysis (PCA)
- Estimation of the kinship coefficient to identify related/duplicated pairs and to check for relatedness consistency.

Form these high quality variants, we are searching for candidate coding variants in patients following an optimal filtering strategy based on: 1) gene-level filtering according to experimental evidence and negative selection level, and: 2) variant-level filtering based on functional annotation, deleteriousness score, frequency in public databases such as gnomAD and mode of inheritance (e.g. autosomal dominant, autosomal recessive or X-linked). We will first search for variants present in patients and not controls that are nonsynonymous, rare (gnomAD MAF<0.01), and potentially deleterious. In multiplex/consanguineous families, we are also performing linkage analysis to narrow down the region of interest. We are conducting a cohort-based analysis by comparing groups of patients with appropriate available SARS-CoV2 infected controls as described in the literature (Zhang 2020, Asano 2021). We are analyzing enrichment at the variant (assuming allelic homogeneity) and gene levels (genetic homogeneity). For common variants, we are comparing the genotype distribution between cases and controls assuming an additive, dominant and recessive model by means of logistic regression and accounting for ethnic heterogeneity of the samples using PCA (Zhang 2020, Asano 2021). For rare variants, we are searching for an enrichment of candidate variants in patients compared with controls using methods to combine variants within a genomic unit for the generation of aggregated statistics as described in (Zhang 2020, Asano 2021). We are also inferring HLA alleles from WES/WGS using HLA*LA (Dilthey 2019) and comparing their frequencies between groups of patients and controls stratified by ethnicity. Analyses are performed according to COVID-19 severity phenotypes and are stratified according to various criteria such as gender, age group, or viral variants.

Results

UNIBO has sent to Whole Genome Sequencing 220 retrospective patients from SAS and UNIBO cohorts: 182 hospitalized patients (166 from SAS; 16 from UNIBO), 38 pauci-symptomatic or a-symptomatic subjects. Clinical data for the hospitalized patients of the SAS cohort are available in 162 patients (97,6%) and are detailed in the table below:

	N°	%
Clinical data	162	97,6
Male	108	65,1
Chronic pulmonary disease	19	11,4
Obesity	33	19,9
Lung infiltrate	149	89,8
Cardiopathy	34	20,5
Hypertension	74	44,6
Asthma	15	9,0
Chronic renal disease	3	1,8
Chronic neurological disease	2	1,2
Diabetes	30	18,1
Chronic inflammatory disease	10	6,0
Smoking	36	21,7
Dislipidemy	34	20,5

INSERM has sequenced 140 severe retrospective patients (63 WES and 77 WGS) from the French Covid cohort. High quality variants have been obtained following the pipelines described in methods. A specific analysis of these patients did not identify any potential deleterious variants in the genes that were previously reported to be involved in severe COViD-19 (Zhang 2020, Asano 2021, Pairo-Castineira 2021). To perform an enrichment analysis in a sufficient number of patients, and to have appropriate SARS-CoV2 infected controls, we merge the data of those patients with previously available samples we have through the COVID Human Genetic effort consortium (Zhang 2020, Asano 2021). We could therefore analyze a total of 1274 severe case and 715 infected controls. Figure 1 shows the ethnic distribution of those subjects into main ethnic groups: African (AFR), North-African (NAFR), European (EUR), Middle-Eastern (ME), South Asian (SAS), American (AMR), and East Asian (EAS). We performed an exome-wide gene-based association study by comparing the proportion of

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

critical cases and infected controls carrying at least one candidate coding variant (CCV) in each gene. We are considering several sets of CCV based on their annotation, predicted impact, and minor allele frequency. Analyses are conducted using a logistic regression model and adjusting for age, gender, and ethnic origin (using the five first principal components). Fig 2 shows the results for a dominant model in the autosomes considering non synonymous variants. No significant associations were observed when correcting for multiple testing. Additional analyses considering alternative models are ongoing.

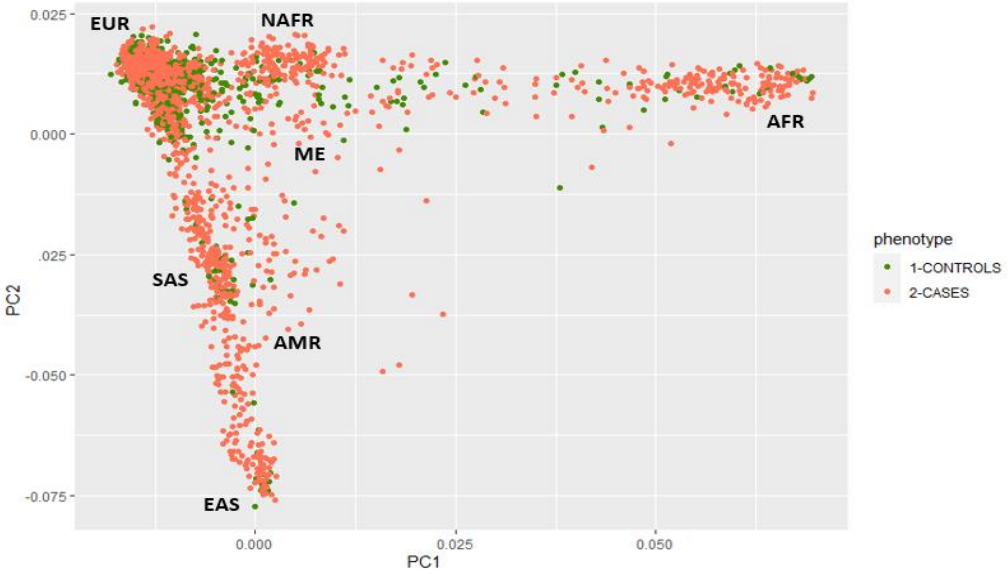

Figure 1: Principal Component Analysis showing the ethnic distribution of patients and controls

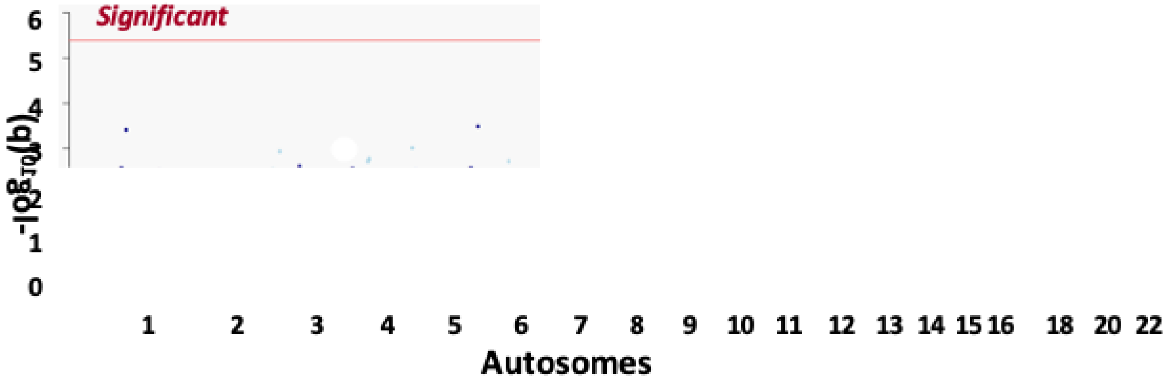

Figure 2: Gene-based enrichment test for rare variants comparing patients with controls using logistic regression adjusted for age, sex, and ethnic origin

Discussion

No significant results for association of severe COVID-19 with rare variants were observed in these preliminary analyses so far. We are now sequencing additional patients to increase the power of the analysis. Some of these patients are also investigated by other collaborators of WP6, in particular for sequencing of the virus (Benoit Visseaux), and we are also testing auto-antibodies against type I IFNs in those samples in order to perform an analysis integrating host and viral genomics as well as immunological parameters as possible risk factors of severe COVID-19.

References

Zhang 2020, DOI: 10.1126/science.abd4570

Asano 2021, doi: 10.1126/sciimmunol.abl4348

Pairo-Castineira 2021, https://doi.org/10.1038/s41586-020-03065-y

Dilthey 2019, doi: 10.1093/bioinformatics/btz235

Acknowledgments

Alessandro Mattiaccio (UNIBO) has worked on samples for WGS/WES from SAS/UNIBO cohorts.