Genomic Determinants of Long COVID

Manuel Corpas (m.corpas@westminster.ac.uk)
University of Westminster

Ilduara Pintos
Puerta de Hierro University Hospital & Research Institute

Víctor Moreno-Torres
Puerta de Hierro University Hospital & Research Institute

Maxim B. Freidin
Queen Mary University London

Segun Fatumo
The African Computational Genomics (TACG) Research Group

Octavio Corral
UNIR Health Sciences School & Medical Center

Vicente Soriano
UNIR Health Sciences School & Medical Center

Carmen Mendoza
Puerta de Hierro University Hospital & Research Institute

Article

**Keywords:** Long COVID, SARS-CoV-2, genetic determinants, low pass whole genome sequencing, polygenic risk scores

**Posted Date:** February 27th, 2023

**DOI:** https://doi.org/10.21203/rs.3.rs-2530935/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.

Read Full License
Abstract

Around 5–10% of adults may experience persistence of symptoms/signs beyond 4 to 12 weeks after acute SARS-CoV-2 infection. According to the World Health Organization, up to 40 million people suffer from Long COVID in Europe and the USA alone. The Centers for Disease Control and Prevention have encouraged the recognition of predictors for Long COVID. Any genetic markers associated to the disease have remained elusive to date. Here we explore the potential contribution of genetic traits to Long COVID. We used a well characterized cohort of 50 individuals with definitive diagnostic criteria for Long COVID from an initial set of patients of more than 1,200 with suspected Long COVID. All were attended at Hospital Puerta de Hierro, a large regional hospital in Madrid, Spain. All subjects had tested positive for SARS-CoV-2 RNA and/or antibodies, showed clinical manifestations for more than 6 months, and developed more than 5 persistent symptoms/signs. Low pass whole genome sequencing was performed in blood specimens for our selected cohort. From hundreds of polygenic risk scores (PRS) recorded at the PGS Catalog, we tested in our selected cohort a total of 12 PRS that passed our filtering criteria. Selected PRS encompassed distinct medical conditions, including cancers, hematologic, cardiovascular, endocrine, immunologic and neurological disorders. The calculated PRS in our patients produced a distribution of scores that was compared to a control ancestry-matched general population. We found significant differences for the PRS of traits ‘Tiredness/lethargy in the last 2 weeks’ and suggestive significance for ‘Depression’ when comparing Long COVID patients and controls. Our results strongly support a genetic susceptibility for Long COVID, with those scoring high in genetic predisposition for ‘tiredness’ as more likely to develop the disease. Results shed new light into the physiopathological basis for Long COVID, contrary to opinions considering it a subjective condition.

Introduction

According to the World Health Organization, around 40 million Long COVID patients have been identified in Europe and the United States alone [1,2]. The Centers for Disease Control and Prevention (CDC), suggests the need to expand the evidence base to estimate both the risk of experiencing post-COVID conditions and the number of people experiencing them by demographic group [3]. Although there are grounds to believe that a genetic signature might be involved in Long COVID [4], to date no genetic markers have been reported. Identification of genetic traits predisposing people to Long COVID could help prevent post-COVID conditions.

The objective of this study is to identify genetic risks for Long COVID based on low coverage whole genome sequencing (LC-WGS). Recently, LC-WGS has been proven a cost-effective way to sequence millions of variants with a reliability of 99% compared to 30x deep coverage by adding a step of imputation [5,6]. LC-WGS has the advantages of wide coverage of variants while generating similar quality variant data using a reference population.

We set out to perform LC-WGS of Long COVID patients aiming at identifying genetic risks based on low coverage plus imputation of a Spanish cohort of Madrid. From a large clinic with access to more than
1,200 patients referred to assess Long COVID, we selected a subset of 50 patients using stringent inclusion criteria (Methods). Their symptoms/signs were annotated and searched for whether they matched existing polygenic risk scores (PRS) from the PGS Catalog, a central repository of PRS containing hundreds of traits [7]. Selected PRS were calculated for each individual patient and controls from the 1000 Genomes Project (1000G) European population [8], comparing differences in the distributions of scores between the patients (cases) and the individuals of the 1000G (controls). We found significant differences between PRS in cases and controls for “Tiredness/lethargy in last 2 weeks” with suggestive significance for “Depression”. This implies a genetic predisposition for Long COVID and the existence of genetic signatures that could be used as a measure of risk for the disease.

Method

Data Cohort

An outpatient clinic for the follow-up of patients with a prior diagnosis of COVID-19 had been set up at one of the largest hospitals in Madrid: Hospital Puerta de Hierro. More than 1,200 patients had been referred with suspected Long COVID for further evaluation at the time our study was conducted. Those with persistent symptoms/signs and with presumed diagnosis of Long COVID were consecutively enrolled in a cohort. For the purpose of this study, strict Long COVID diagnostic criteria were applied. These criteria included prior positive SARS-CoV-2 RNA and/or antibodies, symptoms/signs extending for longer than 6 months after acute infection and more than 5 persistent symptoms/signs, as shown in Table 1. All these clinical manifestations have consistently been recorded in clinical studies focused on Long COVID [9–15], with fatigue as uniformly the most common [16,17].
Table 1
Clinical symptoms/signs used in the diagnosis of Long COVID. Bolded underlined terms that were present in the Polygenic Score Catalog [7]. In our study, we restricted the selection to patients with at least 5 of these symptoms/signs persisting for longer than 6 months after initial testing positive for SARS-CoV-2.

- Fatigue and/or asthenia
- Fever and/or low-grade fever
- Unintentional weight loss / Loss of appetite
- Dizziness and/or unsteadiness
- Insomnia / frequent awakenings / unrefreshing sleep
- “Brain fog”, loss of memory or concentration
- Irritability, emotional lability, anhedonia
- Cough / odynophagia
- Dysphonia / fatigue of the voice
- Dyspnea / choking sensation
- Chest pain
- Palpitations
- Digestive symptoms. Diarrhea, nausea, vomiting, epigastric pain, reflux
- Arthromyalgia, paresthesia, cramps, feeling of numbness of muscles
- Headache / migraine
- Hair loss / Dry skin / brittle nails
- Anosmia / Dysgeusia
- Vision loss
- Hearing loss / tinnitus

We recorded age, sex, country of origin, date of symptoms/signs initiation, and prior comorbidities. For this study we discarded every patient that could have a confounding condition leading to Long COVID-like symptoms, such as fibromyalgia, postraumatic stress syndrome, mental/psychiatric illnesses, chronic neurocognitive diseases, etc. Furthermore, we decided to exclude all patients that had been hospitalized, since recovery from severe illness could have been extended for long periods in a subset of these patients.

**Ethical Framework**
This study was evaluated and approved by the Clinical Research Ethics Committee of Hospital Puerta de Hierro in Madrid, Spain (code number: PI 37/22). In compliance with the provisions of the Declaration of Helsinki and the legislation in force in Spain regarding research with human beings, patients were informed about their participation in the study, clarifying that their participation was voluntary and did not imply any change in his/her treatment or medical care compared to what s/he would receive if s/he did not participate. All patient informed and voluntary consents were obtained in writing.

Sample Handling and Low Pass Whole Genome Sequencing (LP-WGS)

A total of 50 blood samples were collected in 10 mL EDTA tubes. All samples were centrifuged at 3000 rpm during 10 minutes and a buffy coat was isolated and frozen at -20°C until use. Genomic DNA was isolated from buffy coat frozen samples with the Maxwell RSC buffy coat DNA kit (Promega) using the Maxwell RSC Instrument (Promega), following manufacturer's recommendations. After isolation, the purity of genomic DNA was analyzed by spectrometric analysis. Whole genome sequencing was performed using low pass coverage (3 Gigabytes per sample) using MGI. Bioinformatics analysis was performed using mapping and imputation to 1000 Genomes Project using the low impute algorithm [5]. Whole blood was stored for all Long COVID patients as part of a larger study conducted by the hospital’s research unit.

Case / Control Selection

A total of 50 Long COVID patients were selected for cases as described above. One patient failed QC due to contamination by DNA from another sample. None of the patients used were related kin and their ancestries were tested using 1000G populations as reference. As expected, the greatest ancestry detected was Southwestern European, with a minimum of 31% and a mean of 66% among the 49 patients that passed QC. For controls, we used all available genomes within the Iberian Spanish (IBS) subpopulation of the 1000G Phase III (n = 107).

Polygenic Risk Scores (PRS)

PRS were calculated for cases (n = 49) and controls (n = 107) as the sum of the effect weights for all observed risk alleles in a given individual, divided by the total number of risk alleles reported. Scores were independently calculated for all individuals in both case and control populations, each producing a distribution. Three PRS percentiles were collected for each passed QC patient from our imputation provider. These PRS were Coronary Artery Disease [18], Breast Cancer [19], and Prostate Cancer [20]. We report results for these PRS as well.

Apart from these three PRS, which we used as independent control PRS, we searched for all of the symptoms/signs described above to find potential PRS that would be relevant for our Long COVID patients. Apart from matching relevance to Long COVID symptoms/signs, we required that PRS had been published in peer reviewed journals (pre-prints were discarded). All variants in PRS had to record chromosome, position (hg19), effect allele and effect allele weight. We also required that at least 90% of
all SNPs in the PRS were present in our case/control population (i.e., covered by our existing variant data). These criteria reduced our list of PRS to test our patients with to the 9 listed below.

1. Tiredness/lethargy in last 2 weeks (PGS001080) (5594 variants) [21]
2. HDL Cholesterol (PGS000064) (120 variants) [22]
3. LDL Cholesterol (PGS000065) (103 variants) [22]
4. Triglycerides (PGS000066) (101 variants) [22]
5. Venous Thromboembolism (PGS000043) (297 variants) [23]
6. Depression (PGS001829) (7534 variants) [24]
7. Hypothyroidism (PGS000759) (140 variants) [25]
8. Migraine (PGS001281) (25 variants) [21]
9. Asthma (PGS001341) (6430 variants) [21]

**Statistical analysis**

PRS distributions were compared between cases and controls using Wilcoxon rank-sum test (Mann-Whitney test). We tested 9 traits, so the significance threshold was 0.05/9 ~ 0.006.

**Results**

‘Frequency of Tired and Lethargy in the Last Two Weeks’ Scores Significantly Different between Cases and Controls

‘Tiredness/lethargy in the last two weeks’ yielded the strongest differences between cases and controls. Figure 1a shows a scaled density graph comparing scores between cases and controls. Figure 1b shows the same data representing actual scores for cases and controls using boxplots. The observed distributions are significantly different (Wilcoxon rank sum test with continuity correction p-value: 3.43e10^-5).

**PRS for Depression is of suggestive significance in Long COVID patients**

Depression [24] scored a Wilcoxon rank sum test with continuity correction p-value of 0.02813. Because our multiple testing correction reduces our significance threshold to 0.006, we can only classify as being of suggestive significance. As observed in Fig. 2a, the distribution of scores in controls (blue) is only slightly shifted to the left, suggesting a lesser predisposition of controls to suffer from depression than our tested Long COVID patients. Figure 2b shows the same distributions as boxplots with real scores, with medians shown as horizontal lines and scores within 1 standard deviation from the mean at either side.

**No Significant Differences for Other Medical Conditions**
Remaining tested traits included: a) dislipidemias; b) venous thromboembolism; c) endocrine disorders: hypothyroidism; d) neurological diseases: migraine; and e) immunological disorders: asthma. Table 2 provides a summary of all selected PRS tested in our case/control population, with traits in bold that were significantly different in cases and controls.

Table 2: Summary of all tested PRS symptoms in our case / control cohort.

<table>
<thead>
<tr>
<th>Disease Group</th>
<th>Reported Trait</th>
<th>Polygenic Score ID</th>
<th>No. Variants</th>
<th>Citation</th>
<th>Wilcoxon rank sum test with continuity correction p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Systemic</td>
<td>Freq. of tiredness / lethargy in last 2 weeks</td>
<td>PGS001080</td>
<td>5594</td>
<td>Tanigawa Y et al. Plos Genet (2022)</td>
<td>3.43E-05</td>
</tr>
<tr>
<td>Blood</td>
<td>High density lipoprotein (HDL) cholesterol</td>
<td>PGS000064</td>
<td>120</td>
<td>Kuchenbaecker K et al. Nat Commun (2019)</td>
<td>0.7254</td>
</tr>
<tr>
<td></td>
<td>Low density lipoprotein (LDL) cholesterol</td>
<td>PGS000065</td>
<td>103</td>
<td>Kuchenbaecker K et al. Nat Commun (2019)</td>
<td>0.06406</td>
</tr>
<tr>
<td></td>
<td>Triglycerides</td>
<td>PGS000066</td>
<td>101</td>
<td>Kuchenbaecker K et al. Nat Commun (2019)</td>
<td>0.924</td>
</tr>
<tr>
<td>Neurological</td>
<td>Venous thromboembolism</td>
<td>PGS000043</td>
<td>297</td>
<td>Klarin D et al. Nat Genet (2019)</td>
<td>0.4227</td>
</tr>
<tr>
<td>Hormonal</td>
<td>Hypothyroidism</td>
<td>PGS000759</td>
<td>140</td>
<td>Khan Z et al. Nat Commun (2021)</td>
<td>0.9422</td>
</tr>
<tr>
<td>Neurological</td>
<td>Migraine</td>
<td>PGS001281</td>
<td>25</td>
<td>Tanigawa Y et al. Plos Genet (2022)</td>
<td>0.05927</td>
</tr>
<tr>
<td>Immune</td>
<td>Asthma</td>
<td>PGS001341</td>
<td>6430</td>
<td>Tanigawa Y et al. Plos Genet (2022)</td>
<td>0.7717</td>
</tr>
</tbody>
</table>

### Cardiovascular and Cancer PRS for Cases Average ~ 50th Percentile

Coronary Artery Disease (CAD), Breast Cancer (BC) and Prostate Cancer (PC) are among the most common conditions people die of in industrialized countries [26,27]. We collected the risk percentiles for those PRS. Figure 3 shows the average polygenic risk percentile for our 49 Long COVID patients for CAD, BC and PC, showing averages very close to the 50th percentile. We therefore concluded that these traits would not be discriminative or distinctive for Long COVID.

### Discussion

We use low pass whole genome sequencing (LP-WGS) for the analysis of Long COVID patients. LP-WGS’s applications in the clinic have been scant to date, in part due to their novelty [5,6], having little time passed since this technology became widely available. With only two months from sending samples to obtaining meaningful results, LP-WGS allowed us to repurpose genome markers from other studies as useful indicators for genetic susceptibility to Long COVID. The fact that 1000G data are the reference dataset of variant imputation in LP-WGS, makes it possible to cover most variants identified from GWAS analyses and therefore polygenic risk scores (PRS) that make use of GWAS [28]. For a fraction of the cost compared to standard clinical 30x coverage of the whole genome, LP-WGS allows ~ 81 million SNPs from the 1000G to be imputed with 99% accuracy [5]. Our results show the benefit of using a small sample with genome-wide coverage of variants, yet leveraging the value of PRS, which make use of
GWAS involving many thousands of cases in order to assign weights to allele risk contributions to traits tested in validated populations such as the UK Biobank.

With regards to Long COVID, no genomic markers have been published significantly associated to date [29]. This is in part due to lack of definitive datasets that allow clear-cut patients. The definition of clinical symptoms and signs still remains tenuous [4], which makes it difficult to amass datasets sufficiently different and large enough to enable genetic association. Because of these limitations, we have taken a different approach. We make use of existing PRS of likely association with Long COVID symptoms from the PGS Catalog [7]. Our approach of using PRS as a way to test symptoms relevant to Long COVID shows a new way of developing understanding of disease, repurposing existing data for different applications. Our study design thus is able to suggest genetic markers that are highly likely to contribute to Long COVID despite testing a small sample population.

We have been able to show that the PRS trait ‘Tiredness/lethargy in past two weeks’ contains risk allele distributions that are significantly different in Long COVID cases versus controls. We acknowledge the limitation that the control population we use is suboptimal, given that it may also include patients potentially affected by Long COVID, although they would likely be a small fraction. This ‘contamination’ in our control population seems not to impair discrimination between cases vs. controls, suggesting that our stringent inclusion criteria together with using matched general population was an adequate strategy.

We also recognize the limitation that PRS have been trained with Northern European populations, mostly the UK Biobank, which will contain some ancestral genetic differences to the patient cohort in which we tested the PRS with (Iberian Spanish, IBS). In a previous publication [28] we argued that the difference between ancestries within the same continent may influence results but still yield sufficient statistical power for PRS to enable stratification of high risk individuals. The traits or symptoms we have tested for this study do not necessarily overlap 100% with the symptoms described in the clinic (e.g., ‘Tiredness/lethargy in past two weeks’ partially overlaps with fatigue). Yet, the overlap seems significant to infer that a manifestation of fatigue would be concordant with the more specific trait PRS used in our analysis.

Although our case population is predominantly female (~80%), we have not made a distinction between sexes when comparing results, given the small sample size. We leave this question for future work and acknowledge that by including only women in the analysis could slightly change our results. In addition, our approach for selecting symptoms from the PGS Catalog could be interpreted as somewhat ad hoc based on a subjective (although rigorous) selection approach. For future work we plan to test many more PRS following our selection criteria outlined in Methods. We expect that further testing of hundreds of PRS could suggest unsuspected traits and genetic predispositions likely to be associated with Long COVID. We also plan an expanded Long COVID cohort of patients for PRS testing for future work, which could further suggest clues about the pathophysiological mechanisms of the disease.

Our approach does not identify Long COVID-specific markers but PRS with weighted allele risks for independent traits which can be used for stratification of patients with Long COVID genetic
predisposition. In this regard, our methodology is different from other studies that have examined genetic traits associated with tiredness [30], which is the most common symptom reported by Long COVID patients [16,17]. We believe our significantly associated PRS could be utilized as a potential test for predisposition to Long COVID to help predict and prioritize care precisely targeted to the most susceptible patients. Despite the promising results shown here, they will need to be tested in a greater and more diverse population than the one used here.

From results presented here major questions remain. We have not been able to find why Long COVID affects women more than men. We also do not know the exact nature of the association between ‘Tiredness/lethargy in the past two weeks’ and the underlying molecular mechanisms of the disease. It also remains speculation that persons predisposed to depression should be more liable to Long COVID. What our results do suggest, however, is an unequivocal genetic signature for the disease. These results should help dissipate doubts on whether Long COVID is just a subjective condition, where there must be a pathophysiological mechanism behind Long COVID that makes some individuals more prone to develop it.

Finally, our results have been tested on an Iberian Spanish population. However, when comparing cases against controls spanning all 1000G Europeans (n = 503), we get almost identical results (data not shown). This suggests that results are applicable not only to Iberians but Europeans as well. Although it remains speculation, we would also expect similar applications to ancestrally diverse populations although with less precision. Regrettably, on this occasion we only have European patients in our tested cohort. We would like to encourage scientists and patient groups from other ancestries and continents to make it possible to test these hypotheses; so that the fruits of genomic science become accessible to all humanity regardless of ancestral origin.

Conclusions

We make use of low pass whole genome sequencing (LP-WGS) and polygenic risk scores (PRS) to identify genetic determinants for Long COVID in a highly selected cohort of patients. Our results provide strong evidence for association of genetic markers to Long COVID mappable to fatigue. This sheds light on the physiopathological basis for Long COVID, providing evidence against its consideration as a purely subjective condition. Although our results have only been tested in a population of Iberian Spanish patients, we believe they can be extrapolated to other populations and results presented here should serve as the basis for more extensive confirmatory analyses.

Declarations

DATA AVAILABILITY

The sources of PRS used in this study were downloaded from the PGS Catalog (https://www.pgscatalog.org). The datasets generated and/or analyzed during the current study are
available in the EGA repository, with accession number EGAS00001005931.

COMPETING INTERESTS

At the time of this writing, MC is associated with Cambridge Precision Medicine Limited and Genoinsight Limited. No other competing interests have been declared by remaining authors.

FUNDING

Hospital Puerta de Hierro funded this project, including publication costs.

References


5. Li JH, Mazur CA, Berisa T, Pickrell JK. Low-pass sequencing increases the power of GWAS and decreases measurement error of polygenic risk scores compared to genotyping arrays. Genome Res. 2021;31:529–37.


**Figures**

**Figure 1a**

**Figure 1b**

‘Tiredness/lethargy in the last two weeks’ in cases and controls. The scaled density of scores for controls is shown in 1a with blue. Cases are shown in orange. 1b represents boxplots with median as horizontal line and 1 standard deviation as a box. 1 represents controls and 2 represents cases. Both distributions are statistically significantly different (Wilcoxon rank sum test with continuity correction p-value: 3.43e10\(^{-5}\)).
Figure 2

‘Depression’ PRS in cases and controls. The scaled density of scores for controls is shown in 2a with blue. Cases are shown in orange. 2b represents boxplots with median as horizontal line and 1 standard deviation as a box. 1 represents controls and 2 represents cases. Both distributions are of suggestive significance, with a p-value above our multiple testing correction of 0.05/9 ~ 0.006 (Wilcoxon rank sum test with continuity correction p-value: 0.02813).
**Figure 3**

Average polygenic risk percentiles for our 49 Long COVID cases. The y axis shows the percentiles and the x axis the conditions.