mathematics

MDPI

Article

Long-COVID Inducement Mechanism Based on the Path Module Correlation Coefficient

Ziqi Liu [1,2], Ziqiao Yin [2,3,4,5], Zhilong Mi [2,3,4] and Binghui Guo [2,3,4,5,*]

[1] School of Mathematical Sciences, Beihang University, Beijing 100191, China
[2] Key Laboratory of Mathematics, Informatics and Behavioral Semantics and State Key Laboratory of Software Development Environment, Beihang University, Beijing 100191, China
[3] Institute of Artificial Intelligence, Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing, Beihang University, Beijing 100191, China
[4] Peng Cheng Laboratory, Shenzhen 518055, China
[5] Zhongguancun Laboratory, Beijing 100191, China
* Correspondence: guobinghui@buaa.edu.cn

Abstract: As the number of COVID-19 cases increases, the long-COVID symptoms become the focus of clinical attention. Based on the statistical analysis of long-COVID symptoms in European and Chinese populations, this study proposes the path module correlation coefficient, which can estimate the correlation between two modules in a network, to evaluate the correlation between SARS-CoV-2 infection and long-COVID symptoms, providing a theoretical support for analyzing the frequency of long-COVID symptoms in European and Chinese populations. The path module correlation coefficients between specific COVID-19-related genes in the European and Chinese populations and genes that may induce long-COVID symptoms were calculated. The results showed that the path module correlation coefficients were completely consistent with the frequency of long-COVID symptoms in the Chinese population, but slightly different in the European population. Furthermore, the cathepsin C (CTSC) gene was found to be a potential COVID-19-related gene by a path module correlation coefficient correction rate. Our study can help to explore other long-COVID symptoms that have not yet been discovered and provide a new perspective to research this syndrome. Meanwhile, the path module correlation coefficient correction rate can help to find more species-specific genes related to COVID-19 in the future.

Keywords: COVID-19; long COVID symptoms; path module correlation coefficient; path module correlation coefficient correction rate

MSC: 92C42

Citation: Liu, Z.; Yin, Z.; Mi, Z.; Guo, B. Long-COVID Inducement Mechanism Based on the Path Module Correlation Coefficient. *Mathematics* **2023**, *11*, 1368. https://doi.org/10.3390/math11061368

Academic Editors: Sergey A. Lashin and Alfonso Niño

Received: 13 January 2023
Revised: 21 February 2023
Accepted: 9 March 2023
Published: 11 March 2023

1. Introduction

COVID-19 caused by SARS-CoV-2 has greatly affected people's lives. SARS-CoV-2, a coronavirus coated with RNA, shares 79% of its genome sequence with SARS-CoV, which caused the outbreak of SARS in Guangdong, China, in 2002 and 2003, and uses angiotensin-converting enzyme 2 (ACE2) as a receptor into human cells [1]. As of 9 December 2022, the cumulative number of confirmed cases of COVID-19 in the world had frighteningly reached 640 million, with 6.6 million cumulative deaths. Patients infected with COVID-19 usually have fever, sore throat, diarrhea, cough and other symptoms, and the severity of some symptoms is also related to patient age and gender [2]. In addition, infection with COVID-19 may also cause some other concurrent diseases, such as Guillain–Barre syndrome, systemic pneumonia, cerebellar syndrome, etc. [3–5]. Many scholars have researched COVID-19 from several aspects. In terms of COVID-19 pathology, some genes were found to be associated with COVID-19 infection, such as ACE2, HAL-B, DPP4 [6–8]. Dmitry et al. [9] developed a calibrated mathematical model of the antiviral immune response to SARS-CoV-2 infection. They considered multiple immune reaction components in a single calibrated

mathematical model which allowed them to address some fundamental issues related to the pathogenesis of COVID-19. In terms of treatment, hydroxychloroquine and N. sativa seeds maybe helpful to cure COVID-19, but their mechanism of action is not clear [10,11]. In terms of prevention, the locational Hoover index, mean-field models and the susceptible-infected-recovered-deaths model have been used to analyze the transmission and prevention of COVID-19 [12,13]. The Verhulst's, Gompertz's, and SIR models were used to predict the future of the COVID-19 pandemic, using as observed data, the daily cases in the past [14].

Although many people have been cured by treatment after infection, some are not cured in the true sense. A large part of people with a history of COVID-19 infection show various tissue, organ and even systemic symptoms, such as fatigue, shortness of breath, cognitive dysfunction, etc., which affect their daily life [15]. This condition with these symptoms of recovery associated with COVID-19 infection is known as long COVID. Long COVID is characterized by common symptoms present for much longer than expected and lasting effects of the infection in people who have recovered from COVID-19 [16]. Post-acute sequelae of COVID-19, ongoing COVID-19, chronic COVID syndrome, long-haul COVID, and post-COVID-19 are some other terms for long COVID [17]. On 13 September 2022, the WHO reported that modelling data from the Institute for Health Metrics and Evaluation (IHME) at the University of Washington School of Medicine in the United States showed that in 2020 and 2021, nearly 145 million people worldwide had at least one long-COVID symptom. The long-COVID symptoms not only affect normal life, but also can be persistent and recurrent [18]. Although research on COVID-19 is being conducted in depth and has proposed many treatment options, the long-COVID symptoms are not well known [19].

At present, most studies of long-COVID symptoms are based on the statistical data obtained from adults with confirmed SARS-CoV-2 infection [20,21]. The long-COVID symptoms are correlated with the severity of the COVID-19 infection, the time passed since the last vaccination, variants of SARS-CoV-2 and even the gender [22,23]. In Europe, fatigue, dyspnea, joint pain and chest pain are the four symptoms with the highest frequency, and fatigue symptoms account for 53.1% if all symptoms. In contrast, in China, fatigue/muscle weakness is the most common symptom, accounting for 63% of all symptoms, followed by sleeping difficulty and anxiety/depression [24]. In terms of pathophysiology, some specialists studied the long-COVID symptoms by radiology and found lung damage may be responsible for persistent dyspnea and cough in long COVID [25,26]. Some specialists found evidence of the neurotropism and replication capacity of SARS-CoV-2 in neurons in the human brain [27,28]. Since neurons rarely regenerate, the resulting brainstem dysfunction may be long-lasting, leading to neurological and cardiorespiratory sequelae that might underlie long COVID [29]. Some other specialists think SARS-CoV-2 may dysregulate the host immune response, allowing previously harbored pathogens, such as human herpes virus 6 (HHV-6) and human herpes virus 7 (HHV-7), to reactivate, which leads to long COVID [28,30]. In some detailed studies, Russell et al. [31] suggested that the reported long-COVID symptoms can be directly attributed to the dysregulation and chronic activation of cytokine signaling. They found SARS-CoV-2 directly upregulates p38 MAPK genes, potentiates the p38 pathway and blocks p38 MAPK inhibitors, which results in worsening inflammation and disease progression. Rupert [32] proposed that SARS-CoV-2 specifically binds to the membrane Receptor for Advanced Glycation End Products (mRAGE) and Toll Like Receptor 4 (TLR4) and that the pathological positive feedback that sustains the TLR4/RAGE loop is probably the fundamental immunological dysregulation responsible for long COVID. However, the above studies on long COVID were mostly based on retrospective or prospective observational analyses, and theoretical studies of long COVID internal induction mechanisms based on gene regulations are few. COVID-19 has been confirmed to be related to multiple genes, such as ACE2, TMPRSS2, etc. [24,33]. Long-COVID symptoms may be induced by mutation of some genes caused by COVID-19. Therefore, genetic analysis would enable people to find the mechanism supporting the emergence of long-COVID symptoms.

In this work, a path module correlation coefficient based on a human gene interaction network is proposed to measure the correlation between long-COVID symptoms and COVID-19. Firstly, we derived all the gene interaction relationships discovered so far from the NCBI and screened out human gene interaction relationships to obtain the largest connected component (LCC) of the human gene interaction network. Then, according to a previous paper [34], we obtained specific COVID-19-related genes (SCRGs) in the European and Chinese populations and identified the long-COVID symptom-inducing genes (LCSIGs) in Europe and China mentioned in the paper [24] from the NCBI. Then, we calculated the path module correlation coefficient between COVID-19 and each long-COVID symptom based on the path length between SCRGs and LCSIGs and the maximum of shortest paths between nodes within each LCSIGs. We found that the ranks of these correlation coefficients were basically consistent with the frequency of long-COVID symptoms, and only the coefficient of dyspnea for the European population was slightly different. These results provide a theoretical support at the genetic level for the high frequency of long-COVID symptoms in Europe and China. In order to obtain potential European SCRGs, we identified the gene CTSC from Chinese SCRGs through a path module correlation coefficient correction rate. Then, we combined this gene with the European SCRGs to calculate the path module correlation coefficients and showed that the results improved. This suggests that there may be undiscovered European population-specific COVID-19 related genes. This study provides a feasible direction for finding SCRGs in different populations in the future. Figure 1 shows a graphical abstract of this work.

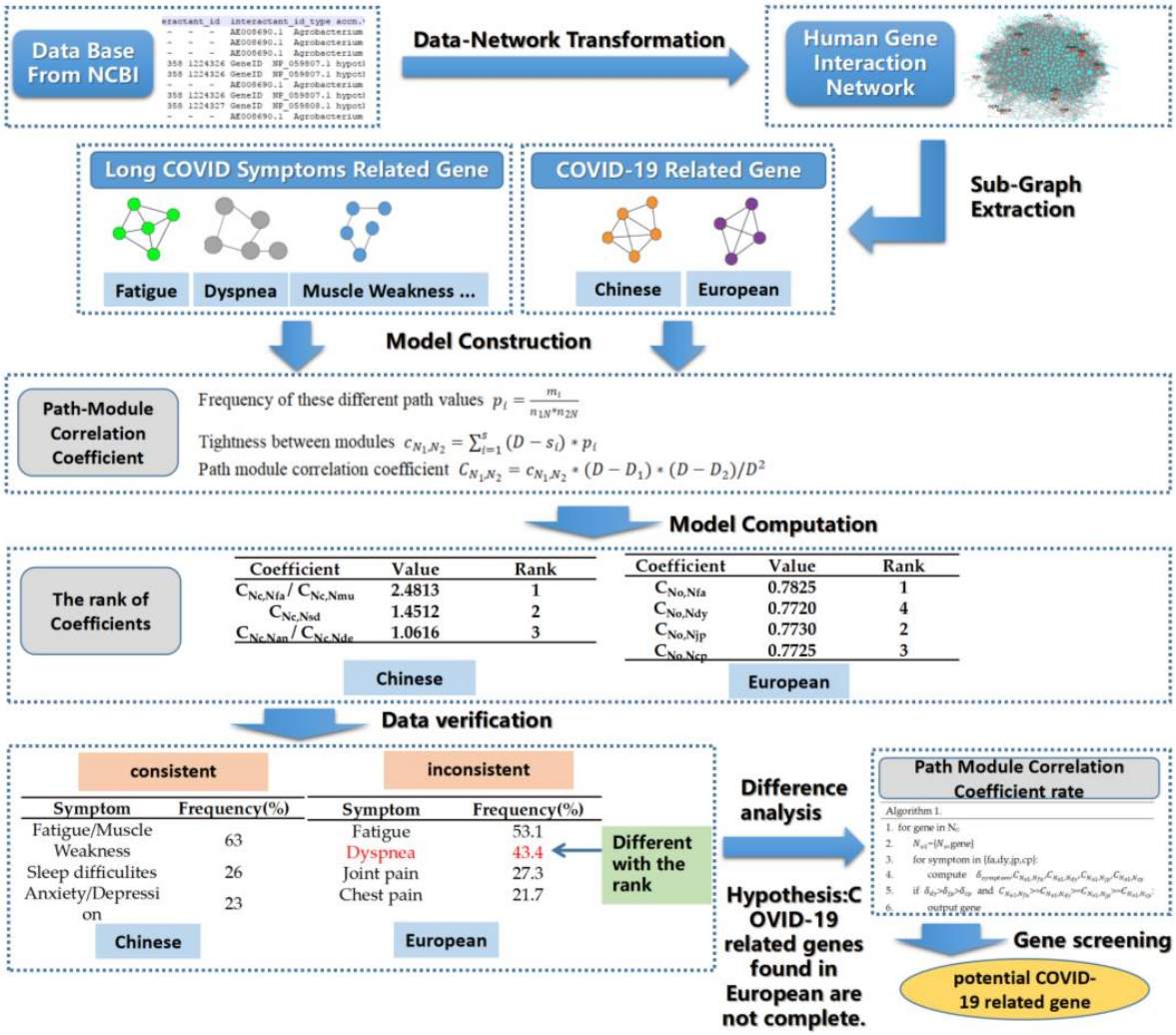


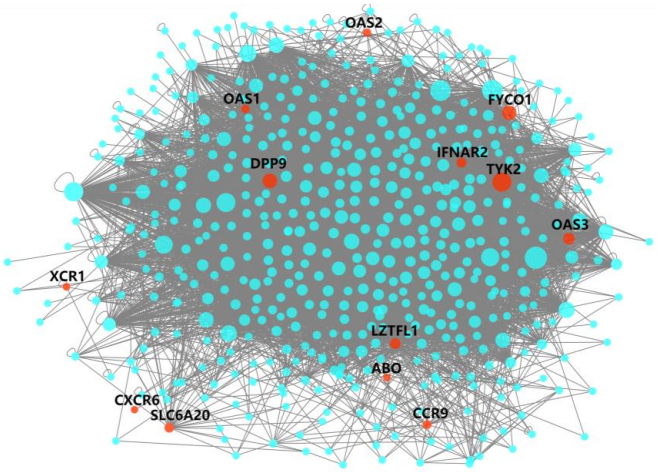
Figure 1. Graphical abstract of this work. The "*" means multiply in the figure.

2. Materials and Methods

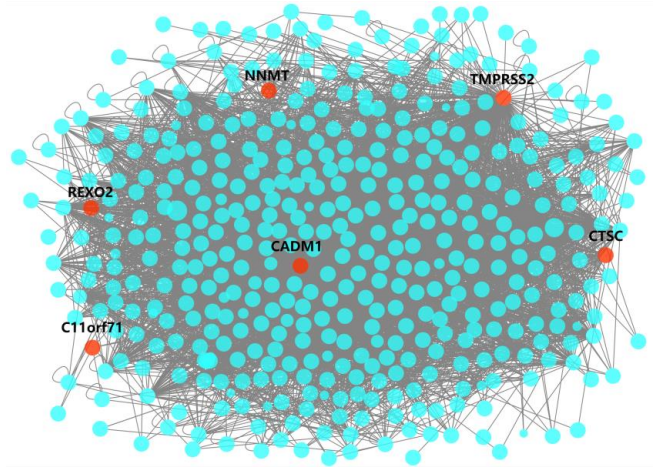
2.1. Data Collection

We obtained the gene interaction relations file from the NCBI database (https://ftp.ncbi.nlm.nih.gov/gene/GeneRIF/interactions.gz, accessed on 17 November 2022). Each line in the file represents an interaction; the information includes the species, gene ID, interaction information, etc., which is supported by the literature. A total of 19,912 genes and 781,952 interactions were obtained by screening human–human gene interactions.

In a previous paper [34], the gene symbols of SCRGs in the European and Chinese populations were provided. We searched the gene IDs corresponding to these genes in the NCBI and matched them with the above-mentioned interaction relationships for subsequent analysis. There were 13 SCRGs in the European population, and 6 SCRGs in the Chinese population. Figure 2 shows the two networks of the SCRGs in the European and Chinese populations with first neighbors. The two networks include 541 nodes and 439 nodes respectively. We can see several hub nodes are present in the SCRGs. Though the SCRGs are not many, the networks of SCRGs with first neighbors are not small. In particular, we did not include in the analysis some widely known COVID-19-related genes such as ACE2, because these genes may exert similar effects in different races, which would not be helpful for our follow-up analysis of long-COVID symptoms in different races.



(**a**)



(**b**)

Figure 2. (**a**) Network of SCRGs in the European population with first neighbors. (**b**) Network of SCRGs in the Chinese population with first neighbors. The red nodes represent SCRGs, and the labels represent the gene symbols. The node size represents the degree of gene. The SCRGs in the European and Chinese populations include several hub nodes.

In a previous paper [24], survey results of long-COVID symptoms in patients in Europe and China were presented. In the survey results for Europe, fatigue (53.1%), dyspnea (43.4%), joint pain (27.3%) and chest pain (21.7%) were the most common symptoms. Fatigue/muscle weakness (63%) was the most common symptom in China, followed by sleeping difficulty (26%) and anxiety/depression (23%). We used symptoms as keywords from NCBI to search for LCSIGs.

2.2. Path Module Correlation Coefficient

In order to measure the correlation between two modules in a complex network, we constructed a path module correlation coefficient based on the path length between nodes in the network. In a connected undirected network G, a path can be found between any two nodes in the network. The node set is defined as $N = \{n_1, n_2, \ldots\ n_N\}$, and the set of edges is defined as $E = \{e_1, e_2 \ldots\ e_E\}$. There were two modules in the network, which means two subsets of node sets, defined as $N_1 = \{n_{11}, n_{12}, \ldots, n_{1N}\}$ and $N_2 = \{n_{21}, n_{22} \ldots\ n_{2N}\}$. The network diameter D is defined as the maximum value of the shortest path between any two nodes in the network, calculated as follows [35]:

$$D = \underset{i,j \in N}{max}\ d_{ij} \tag{1}$$

where d_{ij} represents the shortest path between node i and node j. The path module correlation coefficient of the two network modules N_1 and N_2 is calculated as follows:

1. Take one node from each of the two modules, calculate the shortest path to obtain $n_{1N} * n_{2N}$ shortest paths. These shortest path is $s_1, s_2 \ldots, s_s$, in total, s different path values, and the frequency of these different path values can be calculated as follows:

$$\boldsymbol{p_i} = \frac{\boldsymbol{m_i}}{\boldsymbol{n_{1N}} * \boldsymbol{n_{2N}}}, \boldsymbol{i} = \boldsymbol{1, 2, \ldots, s} \tag{2}$$

where m_i represents the counts of path s_i.

2. Multiply the network diameter minus the shortest path value by the corresponding frequency to obtain the degree of correlation between modules, as follows:

$$\boldsymbol{c_{N_1,N_2}} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{s}} (\boldsymbol{D} - \boldsymbol{s_i}) * \boldsymbol{p_i} \tag{3}$$

3. Calculate the maximum values D_1 and D_2 of the shortest paths between nodes in the two modules. Then, the path module correlation coefficient of the two modules is calculated as follows:

$$\boldsymbol{C_{N_1,N_2}} = \boldsymbol{c_{N_1,N_2}} * (\boldsymbol{D} - \boldsymbol{D_1}) * (\boldsymbol{D} - \boldsymbol{D_2}) / \boldsymbol{D^2} \tag{4}$$

The path module correlation coefficient integrates the degree of proximity between modules and the degree of tightness within modules to obtain the degree of correlation between two modules. The theoretical range of the path module correlation coefficient is $[0, D]$, with 0 indicating that there is no correlation between two modules. The larger the correlation coefficient, the stronger the correlation between the two modules. In fact, the probability that the path module correlation coefficient is 0 or D is very small. When the network is fixed, the path module correlation coefficient of two modules is certain. The Spearman's rank correlation coefficient uses data ranks to calculate the correlation coefficient [36]. Similar to the Spearman's rank correlation coefficient idea, the path module correlation coefficient was used to research the magnitude of correlation between the SCRGs and the LCSIGs and obtain a rank to explain the different frequency of long-COVID symptoms.

2.3. Path Module Correlation Coefficient Correction Rate

If a node is added to the N_1 module, the path module correlation coefficient correction rate is calculated according to the following process:

1. Think of the added node as a separate module N_3. In particular, the maximum shortest path within a module that contains only one node is 0. Calculate the path module correlation coefficient $\mathbf{C_{N_3,N_2}}$ between the modules N_3 and N_2.
2. The path module correlation coefficient correction rate δ after adding nodes is calculated as follows:

$$\delta = \frac{C_{N_3,N_2} - C_{N_1,N_2}}{C_{N_1,N_2}} \tag{5}$$

3. Results

As for the frequency of long-COVID symptoms in the European and Chinese populations previously described [24], we decided to measure the correlation between COVID-19 and different long-COVID symptoms from the perspective of gene networks by calculating the path module correlation coefficients between SCRGs and LCSIGs in Europe and China. Furthermore, a theoretical support is provided by this study for the frequency of long COVID symptoms in different ethnic groups. Considering the inconsistency between the path module correlation coefficient and the frequency, we decided to identify the potential SCRGs targeting specific populations through the path module correlation coefficient correction rate.

3.1. Construction of the Human Gene Interaction Network

We obtained the gene interactions discovered so far from the NCBI database to examine human–human gene interactions and identified 781,952 interactions, involving 19,912 human genes in total. The 19,912 genes were taken as nodes, and the interactions between the genes were taken as edges. An undirected gene interaction network was constructed, in which the LCC contained 19,906 nodes and 781,946 edges. Figure 3 shows the degree distribution of the LCC. It can be seen that the degree distribution of this network met the power law distribution.

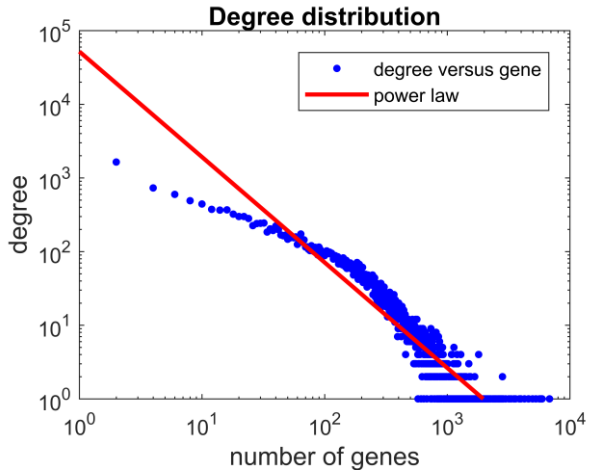


Figure 3. Degree distribution of the LCC. The horizontal coordinate represents the number of genes, the vertical coordinate represents the frequency of the degrees, and the horizontal and vertical coordinates are expressed as double logarithmic coordinates.

3.2. Correlation Analysis between COVID-19 and Long-COVID Symptoms

From a previous paper [34], we obtained SCRGs in the European and Chinese populations, and considered them as the two modules of the LCC, denoting them as N_o and N_c respectively. Genes that may induce fatigue, dyspnea, joint pain, chest pain, muscle weakness, sleep difficulties, anxiety, depression from the NCBI and considered as modules of the LCC, were termed N_{fa}, N_{dy}, N_{jp}, N_{cp}, N_{mu}, N_{sd}, N_{an}, N_{de}, respectively. Table 1 shows the number of genes contained in these modules.

Table 1. Number of genes of each module.

Modules	Number of Genes
N_o	13
N_c	6
N_{fa}	107
N_{dy}	14
N_{jp}	14
N_{cp}	31
N_{mu}	4
N_{sd}	12
N_{an}	260
N_{de}	575

After obtaining the gene modules of COVID-19 and long-COVID symptoms, we calculated the path module correlation coefficients between the SCRGs modules and the LCSIGs modules in the European and Chinese populations. First, the path length and frequency of the SCRGs and LCSIGs were calculated, and the results are presented in Table A1. In Figure 4, we can see that the maximum length of the path between the SCRGs and the LCSIGs in both Europe and China was not more than 5, and in most cases it was 2 or 3, shorter than the network radius, indicating that these long-COVID symptoms are still closely related to COVID-19 at the genetic level. Then, we calculated the maximum value of the shortest path between genes in the SCRGs module and genes in the LCSIGs module. According to the results in Table 2, we can see that the maximum value of the shortest path in most modules was 3 or 4, and only the value of the anxiety and depression gene modules reached 5 and 6, indicating that the gene connection in most modules was also relatively strong, while the results for the anxiety and depression gene modules could be due to the large number of genes contained in these modules.

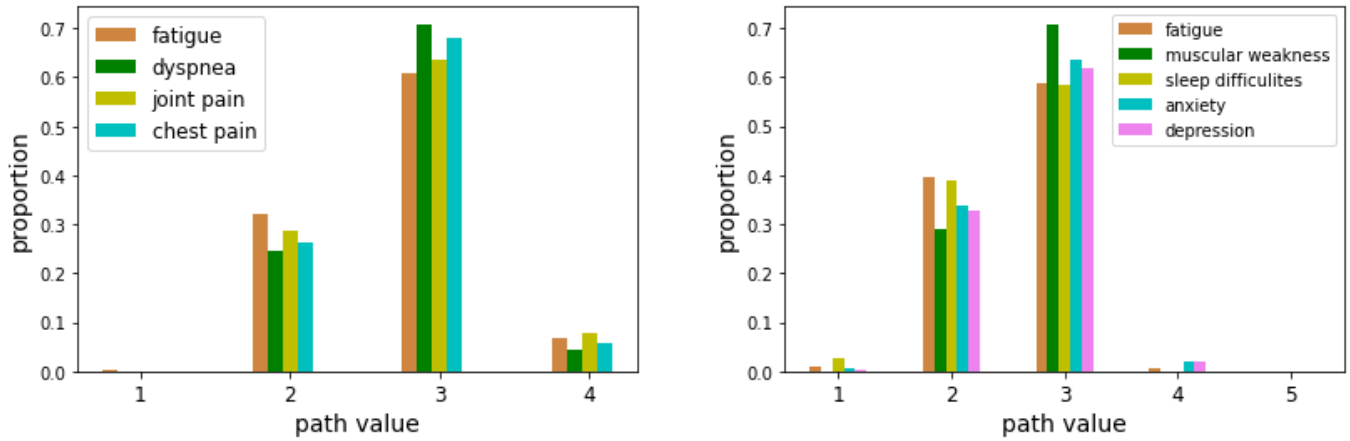


Figure 4. (**a**) Path values and proportions of COVID-19 genes and different symptom genes in Europe. (**b**) Path values and frequencies of COVID-19 genes and different symptom genes in China. The x-axis represents the different path values, the y-axis represents the proportion of the path values. The maximum length of the path between the SCRGs and the LCSIGs in both Europe and China was not more than 5, and most path values were 2 and 3, indicating that these long-COVID symptoms are closely related to COVID-19 at the genetic level.

Table 2. Maximum value of the shortest path in each module. The maximum value of the shortest paths in most modules was 3 or 4, and only the value of the anxiety and depression gene modules reached 5 and 6, indicating that the gene connection in most modules was relatively strong.

Modules	Maximum Value of the Shortest Path
N_o	4
N_c	3
N_{fa}	4
N_{dy}	4
N_{jp}	4
N_{cp}	4
N_{mu}	3
N_{sd}	3
N_{an}	5
N_{de}	6

Finally, according to steps 2 and 3 in Section 2.2, the path module correlation coefficients between the European and the Chinese SCRGs modules and the LCSIGs modules were calculated. Table 3a shows the path module correlation coefficients between the European SCRGs modules and the corresponding LCSIGs modules, and Table 3b shows the Chinese results. We intercepted four decimal places, which is enough to obtain an accurate rank of the path module correlation coefficients. The final columns in Tables 3a,b report some statistical data about the frequency of long-COVID symptoms, as previously indicated [24]. Although the values in Table 3a are close, their small difference may become a large correlation difference for a large network; meanwhile, our focus was on the rank. It was found for the European results, that the rank of the path module correlation coefficient calculated by us is basically consistent with the rank of the long-COVID symptoms frequency previously described [24], except for the symptom of dyspnea. The correlation coefficient between fatigue symptom genes and COVID-19 genes was the largest. Previous research [24] concluded that the frequency of fatigue symptoms in the European population was the highest. The second correlation coefficient was that for the correlation between joint pain symptom genes and COVID-19 genes, followed by the correlation coefficient between chest pain symptom genes and COVID-19 genes. As for the results of the Chinese population, the frequencies of fatigue and muscle weakness symptoms, as well as those of anxiety and depression symptoms were combined in the paper [24]. The path module correlation coefficients of combined fatigue and muscle weakness symptoms genes and COVID-19 genes were determined, as well as those of combined anxiety and depression symptoms genes and COVID-19 genes. The results showed the path module correlation coefficient of fatigue/muscle weakness and COVID-19 was greater than that of sleep difficulty symptoms and anxiety/depression symptoms, which is completely consistent with the rank of long-COVID symptoms' frequency described in the paper [24]. Therefore, according to our results, the path module correlation coefficient proposed by us indicates a long-COVID inducement mechanism operating at the gene level.

As for the path module correlation coefficient value, we performed some experiments. We randomly selected two modules, each including 10 nodes, from the gene interaction network and calculated the path module correlation coefficient between them. Then, two modules including 50, 100, 150, 200 nodes were randomly selected for the same operation. Each experiment was repeated fifty times. Table 4 shows the results of the distribution frequency of the path module correlation coefficients. We can see that all of the coefficients fell into the [0.1, 1.5) interval, most coefficients fell into the [0.3, 0.9) interval, and a small number of coefficients fell into the [0.9, 1.5) interval. Therefore, the path module correlation coefficients we calculated are reasonable.

Table 3. (**a**). Path module correlation coefficient between the European SCRGs module and the corresponding LCSIGs module. The rank of the path module correlation coefficient calculated by us is basically consistent with the rank of the long-COVID symptoms frequency described in the paper [24], except for the symptom of dyspnea. (**b**). Path module association coefficient between the Chinese SCRGs module and the corresponding LCSIGs module. The path module correlation coefficients of combined fatigue and muscle weakness symptoms genes and the COVID-19 genes and combined anxiety and depression symptoms genes and the COVID-19 genes were determined. The rank of the path module correlation coefficients is completely consistent with the rank of the long-COVID symptoms' frequency described in the paper [24].

(a)			
Coefficient	**Value**	**Rank**	**Symptom Frequency (%)**
$C_{No,Nfa}$	0.7825	1	53.1
$C_{No,Ndy}$	0.7720	4	43.4
$C_{No,Njp}$	0.7730	2	27.3
$C_{No,Ncp}$	0.7725	3	21.7
(b)			
Coefficient	**Value**	**Rank**	**Symptom Frequency (%)**
$C_{Nc,Nfa}/C_{Nc,Nmu}$	2.4813	1	63
$C_{Nc,Nsd}$	1.4512	2	26
$C_{Nc,Nan}/C_{Nc,Nde}$	1.0616	3	23

Table 4. Distribution frequency of the path module correlation coefficients between two modules including 10, 50, 100, 150, 200 nodes. All of the coefficients fall into the [0.1, 1.5) interval, most coefficients fell into the [0.3, 0.9) interval, and a small number of coefficients fell into the [0.9, 1.5) interval.

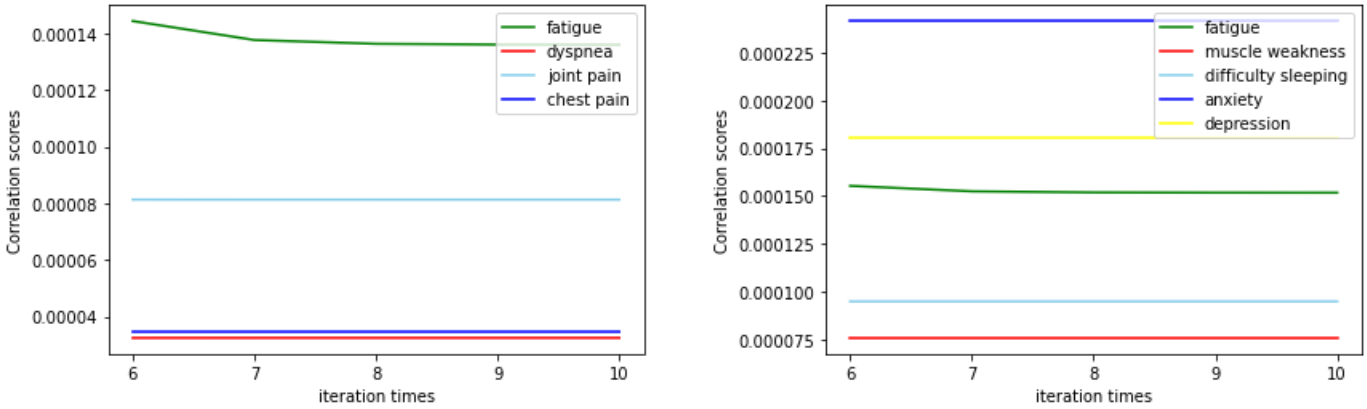
Distribution Interval	**Frequency (%)**
[0, 0.1)	0
[0.1, 0.3)	4.8
[0.3, 0.5)	42.4
[0.5, 0.7)	23.6
[0.7, 0.9)	17.2
[0.9, 1.1)	8.8
[1.1, 1.3)	0
[1.3, 1.5)	3.2

For comparison, we also used a random walk algorithm with restart to evaluate the association between long-COVID symptoms and COVID-19. The random walk with restart is defined by the following Equation [37]:

$$\vec{r}_l = c\widetilde{W}\vec{r}_l + (1-c)\vec{e}_l \tag{6}$$

where $\vec{r}_l$ is the correlation score between the start node i and the other nodes, $\widetilde{W}$ represents the standardization of the network connection matrix, c is the probability of restart, $\vec{e}_l$ is a unit vector of $n * 1$, the ith element is 1, the remaining elements are 0, n is the number of nodes in the network. For a long-COVID symptom, we took each symptom gene as the starting point of migration, calculated the correlation scores between the symptom genes and the COVID-19 genes, and calculated the average value as the correlation score between the long-COVID symptom and COVID-19. Due to the large number of network nodes, we adopted an iterative method to solve the problem. As can be seen in the Figure 5, the correlation scores tended to converge at the 9th step. For a clear comparison, Figure 5 shows only the results of the last five steps. We can see that the rank of the correlation scores between COVID-19 and long-COVID symptoms in both Europe and China obtained by the

random walk algorithm with restart were inconsistent with the frequency of symptoms reported in the paper [24]. The final correlation scores are presented in Table A2. We further demonstrated that the path module correlation coefficient proposed by us has theoretical support for determining the occurrence probability of long-COVID symptoms.



(**a**) (**b**)

Figure 5. (**a**) Correlation scores between COVID-19 and long-COVID symptoms in Europe. (**b**) Correlation scores between COVID-19 and long-COVID symptoms in China. The x-axis represents the iteration times, and the y-axis represents the correlation scores. The rank of the correlation scores between COVID-19 and long-COVID symptoms in both Europe and China obtained by the random walk algorithm with restart are inconsistent with the frequency of symptoms reported in the paper [24].

3.3. Potential European Population-Specific COVID-19-Related Genes

As can be seen from the results in Table 3a, the path module correlation coefficient between the dyspnea symptom gene and the COVID-19 genes for the European population was smaller than that of joint pain and chest pain, which is inconsistent with the frequency of dyspnea symptoms. The reason may be the specific COVID-19-related genes found in the European population were not complete. We therefore mined potential European SCRGs based on a path module correlation coefficient correction rate. Based on the discovered SCRGs in the Chinese population, we explored whether there are SCRGs targeting the European population. Algorithm 1 is as follows:

Algorithm 1

1. for gene in N_c:
2. N_{o1} = {N_o,gene}
3. for symptom in {fa,dy,jp,cp}:
4. compute $\delta_{symptom}$,$C_{N_{o1},N_{fa}}$,$C_{N_{o1},N_{dy}}$,$C_{N_{o1},N_{jp}}$,$C_{N_{o1},N_{cp}}$
5. if $\delta_{dy} > \delta_{jp} > \delta_{cp}$ and $C_{N_{o1},N_{fa}} \geq C_{N_{o1},N_{dy}} \geq C_{N_{o1},N_{jp}} \geq C_{N_{o1},N_{cp}}$:
6. output gene

According to Algorithm 1, the gene CTSC was obtained. Table 5 shows the path module correlation coefficient between the new module N_{o1}, which contained the gene CTSC, and the gene modules of fatigue, dyspnea and other symptoms. Although the values did not change much from Table 3a, in terms of rank, the path module correlation coefficients between N_{o1} and the long-COVID symptom gene modules were more consistent with the frequency of the long-COVID symptoms than before, indicating that CTSC may be a potential European SCRG. Of course, there may be other potential genes to be discovered. This also provides a feasible direction for the future search for SCRGs in different populations.

Table 5. Path module correlation coefficients between the new module N_{o1} and the modules N_{fa},N_{dy},N_{jp} and N_{cp}. Compared with the results in Table 3a, the results in this table are more in line with the reality.

Coefficient	Value	Rank
$C_{No1,Nfa}$	0.784	1
$C_{No1,Ndy}$	0.774	2
$C_{No1,Njp}$	0.774	2
$C_{No1,Ncp}$	0.773	3

4. Discussion

According to the results of IHME, nearly 145 million people have experienced long COVID; therefore, research on the inducement mechanism of long COVID and the development of effective prevention and therapeutic measures have become an urgent problem. Fatigue, dyspnea, cough, headache, brain fog, loss of smell and taste disorders are common long-COVID symptoms. Many studies on long-COVID symptoms were conducted only to obtain some statistical data from symptom reports or return visits of patients and did not considered in depth the internal correlation between these symptoms and COVID-19 from the perspective of genes [38,39]. Therefore, we hoped to obtain such internal correlation by analyzing human gene networks.

Traditional correlation coefficients such as the Pearson correlation coefficient, Spearman correlation coefficient and Kendall correlation coefficient are often used to measure the correlation of two variables X and Y. In order to measure the correlation between two modules in a complex network, the tightness of nodes within modules and the tightness of nodes between modules should be considered together, and traditional correlation coefficients seem useless. The tightness of nodes can be estimated by the path length; so, we construct a path module correlation coefficient based on the path length between nodes in the network. In this study, the correlation between COVID-19 and long-COVID symptoms was evaluated by the path module correlation coefficient at the gene level. The path module correlation coefficient combined the internal and inter-module tightness of gene modules related to COVID-19 that may induce long-COVID symptoms. We calculated the path module correlation coefficient between the SCRGs modules and the LCSIGs modules in European and Chinese populations. The results showed that the rank of the path module correlation coefficients was consistent with the rank of the frequency of long-COVID symptoms in Chinese patients as well as in European patients, except for dyspnea, which indicated that the path module correlation coefficient proposed by us has a certain theoretical validity for determining the frequency of long-COVID symptoms. Moreover, the results also showed that the fatigue symptom had different path module correlation coefficients calculated in relation to SCRGs in the two races, indicating that the occurrence of long-COVID symptoms may be related to race.

We considered that the path module correlation coefficients for Europe were not completely consistent with the rank of the long-COVID symptoms' frequency in European patients, possibly due to the incomplete discovery of SCRGs in the European population. Because of the validity of the SCRGs for the Chinese population, we tried to understand whether a gene can be added to the SCRGs for the European population, so that the path module correlation coefficients between the new SCRGs modules in the European population and the LSCRGs modules of the corresponding long-COVID symptoms become more representative of the real situation. We screened the desired target gene based on the path module correlation coefficient correction rate, and the result showed that the gene CTSC satisfied our expectations. Although it could not be determined whether the gene is an SCRG in the European population, the addition of this gene could make the rank of the path module correlation coefficients more close to the actual situation, suggesting the presence of potential European population-specific COVID-19-related genes.

Our research provides a mechanism for the emergence of long COVID. In the diagnosis and treatment of long COVID, the path module correlation coefficient can help to estimate whether a disease symptom is a long-COVID symptom and bring it to clinical attention. Meanwhile, the path module correlation coefficient can be also used in precision medicine. Doctors can diagnose this syndrome in individual patient based on unique gene sequencing data. The path module correlation coefficient correction rate can help to explore the potential SCRGs that are strongly related to the LSCRGs according to simple and easily available representation data, such as the frequency of long-COVID symptoms. In future work, a promising research line is to further investigate long-COVID symptoms to identify still unknown ones by the path module correlation coefficient, so that some preventive measure can be taken. Furthermore, our work can be extended to construct more efficient models and find more potential COVID-19-related genes in the entire range of human genes.

Author Contributions: Data curation, Z.M.; Methodology, Z.Y.; Visualization, Z.M.; Writing—original draft, Z.L.; Writing—review & editing, B.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key R&D Program of China (Grant: 2021ZD0201302), the Key R&D Program of Guangdong Province, China (Grant: 2021B0101420003), the National Natural Science Foundation of China (Grant: 12201025), the National Natural Science Foundation of China (Grant: U20B2053).

Data Availability Statement: The experiment data are available at https://ftp.ncbi.nlm.nih.gov/gene/GeneRIF/interactions.gz, accessed on 17 November 2022.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. (**a**). Path values and proportions of COVID-19 genes and different symptom genes in Europe. (**b**). Path values and frequencies of COVID-19 genes and different symptom genes in China.

(a)				
	1	**2**	**3**	**4**
fatigue	6/1391	445/1391	845/1391	95/1391
dyspnea	0	45/182	129/182	8/182
joint pain	0	52/182	116/182	14/182
chest pain	0	106/403	274/403	23/403

(**b**)					
	1	**2**	**3**	**4**	**5**
fatigue	6/642	254/642	378/642	4/642	0
muscle weakness	0	7/24	17/24	0	0
sleep difficulties	2/72	28/72	42/72	0	0
anxiety	9/1560	527/1560	994/1560	30/1560	0
depression	13/3450	1132/3450	2129/3450	74/3450	3/3450

Table A2. (**a**). Correlation scores between COVID-19 and long-COVID symptoms in Europe. (**b**). Correlation scores between COVID-19 and long-COVID symptoms in China.

(**a**)	
Correlation Scores	**Value**
$r_{o\text{-}fa}$	0.000136
$r_{o\text{-}dy}$	0.000033
$r_{o\text{-}jp}$	0.000081
$r_{o\text{-}cp}$	0.000035
(**b**)	
Coefficient	**Value**
$r_{c\text{-}fa}$	0.000152
$r_{c\text{-}mu}$	0.000076
$r_{c\text{-}sd}$	0.000095
$r_{c\text{-}an}$	0.000241
$r_{c\text{-}de}$	0.000181

References

1. Kenny, G.; Mallon, P.W. COVID19- clinical presentation and therapeutic considerations. *Biochem. Biophys. Res. Commun.* **2021**, *538*, 125–131. [CrossRef] [PubMed]
2. Elibol, E. Otolaryngological symptoms in COVID-19. *Eur. Arch. Otorhinolaryngol.* **2021**, *278*, 1233–1236. [CrossRef] [PubMed]
3. Aladawi, M.; Elfil, M.; Abu-Esheh, B.; Jazar, D.A.; Armouti, A.; Bayoumi, A.; Piccione, E. Guillain Barre Syndrome as a Complication of COVID-19: A Systematic Review. *Can. J. Neurol. Sci.* **2022**, *49*, 38–48. [CrossRef] [PubMed]
4. de Oliveira Filho, C.M.; Vieceli, T.; de Fraga Bassotto, C.; da Rosa Barbato, J.P.; Garcia, T.S.; Scheffel, R.S. Organizing pneumonia: A late phase complication of COVID-19 responding dramatically to corticosteroids. *Braz. J. Infect. Dis.* **2021**, *25*, 101541. [CrossRef]
5. Šín, R.; Sedláček, D.; Cmorej, P.C.; Peřan, D. Cerebellar syndrome as a complication of COVID-19 disease. *Neuroendocrinol. Lett.* **2021**, *42*, 508–511.
6. Carter, C.; Airas, J.; Parish, C.A. Atomistic insights into the binding of SARS-CoV-2 spike receptor binding domain with the human ACE2 receptor: The importance of residue 493. *J. Mol. Graph. Model.* **2023**, *118*, 108360. [CrossRef]
7. Zhang, Y.; Sun, Y.; Zhu, H.; Hong, H.; Jiang, J.; Yao, P.; Liao, H.; Zhang, Y. Allelic imbalance of HLA-B expression in human lung cells infected with coronavirus and other respiratory viruses. *Eur. J. Hum. Genet.* **2022**, *30*, 922–929. [CrossRef]
8. Nádasdi, Á.; Sinkovits, G.; Bobek, I.; Lakatos, B.; Förhécz, Z.; Prohászka, Z.; Réti, M.; Arató, M.; Cseh, G.; Masszi, T.; et al. Decreased circulating dipeptidyl peptidase-4 enzyme activity is prognostic for severe outcomes in COVID-19 inpatients. *Bio. Med.* **2022**, *16*, 317–330. [CrossRef]
9. Grebennikov, D.; Karsonova, A.; Loguinova, M.; Casella, V.; Meyerhans, A.; Bocharov, G. Predicting the Kinetic Coordination of Immune Response Dynamics in SARS-CoV-2 Infection: Implications for Disease Pathogenesis. *Mathematics* **2022**, *10*, 3154. [CrossRef]
10. Sarma, P.; Kaur, H.; Kumar, H.; Mahendru, D.; Avti, P.; Bhattacharyya, A.; Prajapat, M.; Shekhar, N.; Kumar, S.; Singh, R.; et al. Virological and clinical cure in COVID-19 patients treated with hydroxychloroquine: A systematic review and meta-analysis. *J. Med. Virol.* **2020**, *92*, 776–785. [CrossRef]
11. Islam, M.N.; Hossain, K.S.; Sarker, P.P.; Ferdous, J.; Hannan, A.; Rahman, M.; Chu, D.-T.; Uddin, J. Revisiting pharmacological potentials of Nigella sativa seed: A promising option for COVID-19 prevention and cure. *Phytot. Res.* **2021**, *35*, 1329–1344. [CrossRef] [PubMed]
12. Fanelli, D.; Piazza, F. Analysis and forecast of COVID-19 spreading in China, Italy and France. *Chaos Solitons Fractals* **2020**, *134*, 109761. [CrossRef]
13. Castro, M.C.; Kim, S.; Barberia, L.; Ribeiro, A.F.; Gurzenda, S.; Ribeiro, K.B.; Abbott, E.; Blossom, J.; Rache, B.; Singer, B.H. Spatiotemporal pattern of COVID-19 spread in Brazil. *Science* **2021**, *372*, 821–826. [CrossRef]
14. Martínez-Fernández, P.; Fernández-Muñiz, Z.; Cernea, A.; Fernández-Martínez, J.L.; Kloczkowski, A. Three Mathematical Models for COVID-19 Prediction. *Mathematics* **2023**, *11*, 506. [CrossRef]
15. Soriano, J.B.; Murthy, S.; Marshall, J.C.; Relan, P.; Diaz, J.V. A clinical case definition of post-COVID-19 condition by a Delphi consensus. *Lancet Infect. Dis.* **2022**, *22*, e102–e107. [CrossRef] [PubMed]
16. Mahase, E. Covid-19: What do we know about "long covid"? *BMJ* **2020**, *370*, m2815. [CrossRef]
17. O'Mahoney, L.L.; Routen, A.; Gillies, C.; Ekezie, W.; Welford, A.; Zhang, A.; Karamchandani, U.; Simms-Williams, N.; Cassambai, S.; Ardavani, A.; et al. The prevalence and long-term health effects of Long Covid among hospitalised and non-hospitalised populations: A systematic review and meta-analysis. *eClinicalMedicine* **2023**, *55*, 101762. [CrossRef] [PubMed]
18. Raveendran, A.V.; Jayadevan, R.; Sashidharan, S. Long COVID: An overview. *Diabetes Metab. Syndr. Clin. Res. Rev.* **2021**, *15*, 869–875. [CrossRef]

19. Lu, Y.; Chen, L.; Gu, J.-N.; Yang, X.-Y. Advances in "Long COVID" syndrome. *Chin. J. Nosocomiol.* **2022**, *32*, 2556–2560.
20. Subramanian, A.; Nirantharakumar, K.; Hughes, S.; Myles, P.; Williams, T.; Gokhale, K.M.; Taverner, T.; Chandan, J.S.; Brown, K.; Simms-Williams, N.; et al. Symptoms and risk factors for long COVID in non-hospitalized adults. *Nat. Med.* **2022**, *28*, 1706–1714. [CrossRef]
21. Thompson, E.J.; Williams, D.M.; Walker, A.J.; Mitchell, R.E.; Niedzwiedz, C.L.; Yang, T.C.; Huggins, C.F.; Kwong, A.S.F.; Silverwood, R.J.; Di Gessa, G.; et al. Long COVID burden and risk factors in 10 UK longitudinal studies and electronic health records. *Nat. Commun.* **2022**, *13*, 3528. [CrossRef]
22. Global Burden of Disease Long COVID Collaborators. Estimated Global Proportions of Individuals with Persistent Fatigue, Cognitive, and Respiratory Symptom Clusters Following Symptomatic COVID-19 in 2020 and 2021. *JAMA* **2022**, *328*, 1604–1615. [CrossRef] [PubMed]
23. Antonelli, M.; Pujol, J.C.; Spector, T.D.; Ourselin, S.; Steves, C.J. Risk of long COVID associated with delta versus omicron variants of SARS-CoV-2. *Lancet* **2022**, *399*, 2263–2264. [CrossRef]
24. Nalbandian, A.; Sehgal, K.; Gupta, A.; Madhavan, M.V.; McGroder, C.; Stevens, J.S.; Cook, J.R.; Nordvig, A.S.; Shalev, D.; Sehrawat, T.S.; et al. Post-acute COVID-19 syndrome. *Nat. Med.* **2021**, *27*, 601–615. [CrossRef]
25. Li, H.; Zhao, X.; Wang, Y.; Lou, X.; Chen, S.; Deng, H.; Shi, L.; Xie, J.; Tang, D.; Zhao, J.; et al. Damaged lung gas-exchange function of discharged COVID-19 patients detected by hyperpolarized (129)Xe MRI. *Sci. Adv.* **2020**, *7*, eabc8180. [CrossRef] [PubMed]
26. Swigris, J.J.; Streiner, D.L.; Brown, K.K.; Belkin, A.; Green, K.E.; Wamboldt, F.S.; IPFnet Investigators. Assessing exertional dyspnea in patients with idiopathic pulmonary fibrosis. *Respir. Med.* **2014**, *108*, 181–188. [CrossRef] [PubMed]
27. Zhang, B.; Chu, H.; Han, S.; Shuai, H.; Deng, J.; Hu, Y.; Gong, H.; Lee, A.; Zou, Z.; Yau, T.; et al. SARS-CoV-2 infects human neural progenitor cells and brain organoids. *Cell Res.* **2020**, *30*, 928–931. [CrossRef]
28. Proal, A.D.; VanElzakker, M.B. Long COVID or Post-acute Sequelae of COVID-19 (PASC): An Overview of Biological Factors That May Contribute to Persistent Symptoms. *Front. Microbiol.* **2021**, *12*, 698169. [CrossRef]
29. Shin, J.y. Persistent Brainstem Dysfunction in Long-COVID: A Hypothesis. *ACS Chem. Neurosci.* **2021**, *12*, 573–580.
30. Drago, F.; Ciccarese, G.; Rebora, A.; Parodi, A. Human herpesvirus-6, -7, and epstein-barr virus reactivation in pityriasis rosea during COVID-19. *J. Med. Virol.* **2021**, *93*, 1850–1851. [CrossRef]
31. Low, R.N.; Low, R.J.; Akrami, A. A Cytokine-based model for the pathophysiology of Long COVID symptoms. *OSF Prepr.* **2020**, *10*, 1473–1487.
32. Holms, R.D. Long COVID (PASC) Is Maintained by a Self-Sustaining Pro-Inflammatory TLR4/RAGE-Loop of S100A8/A9 > TLR4/RAGE Signalling, Inducing Chronic Expression of IL-1b, IL-6 and TNFa: Anti-Inflammatory Ezrin Peptides as Potential Therapy. *Immuno* **2022**, *2*, 512–533. [CrossRef]
33. Kuba, K.; Imai, Y.; Rao, S.; Shuai, H.; Deng, J.; Hu, Y.-F.; Gong, H.-R.; Lee, A.C.-Y.; Zou, Z.; Yau, T.; et al. A crucial role of angiotensin converting enzyme 2 (ACE2) in SARS coronavirus–induced lung injury. *Nat. Med.* **2005**, *11*, 875–879. [CrossRef]
34. Li, Y.; Ke, Y.; Xia, X.; Wang, Y.; Cheng, F.; Liu, X.; Jin, X.; Li, B.; Xie, C.; Liu, S.; et al. Genome-wide association study of COVID-19 severity among the Chinese population. *Cell Discov.* **2021**, *7*, 76. [CrossRef] [PubMed]
35. Kleinberg, J. Complex networks and decentralized search algorithms. In Proceedings of the International Congress of Mathematicians, Hong Kong, China, 18–22 December 2006.
36. Gauthier, T.D. Detecting Trends Using Spearman's Rank Correlation Coefficient. *Environ. Forensics* **2001**, *2*, 359–362. [CrossRef]
37. Tong, H.; Faloutsos, C.; Pan, J.-Y. Fast Random Walk with Restart and Its Applications. In Proceedings of the Sixth International Conference on Data Mining, Washington, DC, USA, 18–22 December 2006.
38. Buonsenso, D.; Munblit, D.; De Rose, C.; Sinatti, D.; Ricchiuto, A.; Carfi, A.; Valentini, P. Preliminary evidence on long COVID in children. *Acta Paediatr.* **2021**, *110*, 2208–2211. [CrossRef]
39. Fernández-de-las-Peñas, C.; Palacios-Ceña, D.; Gómez-Mayordomo, V.; Cuadrado, M.L.; Florencio, L.L. Defining Post-COVID Symptoms (Post-Acute COVID, Long COVID, Persistent Post-COVID): An Integrative Classification. *Int. J. Environ. Res. Public Health* **2021**, *18*, 2621. [CrossRef]