Briefings in Bioinformatics, 00(00), 2021, 1–16

doi: 10.1093/bib/bbab169
Problem Solving Protocol

The peripheral and core regions of virus-host network of COVID-19

Bingbo Wang, Xianan Dong, Jie Hu, Xiujuan Ma, Chao Han, Yajun Wang and Lin Gao

Corresponding author: Bingbo Wang, School of Computer Science and Technology, Xidian University, Xi'an 710071, China. Tel.: 86-29-88202354; E-mail: bingbowang@xidian.edu.cn

Abstract

Two thousand nineteen novel coronavirus SARS-CoV-2, the pathogen of COVID-19, has caused a catastrophic pandemic, which has a profound and widespread impact on human lives and social economy globally. However, the molecular perturbations induced by the SARS-CoV-2 infection remain unknown. In this paper, from the perspective of omnigenic, we analyze the properties of the neighborhood perturbed by SARS-CoV-2 in the human interactome and disclose the peripheral and core regions of virus-host network (VHN). We find that the virus-host proteins (VHPs) form a significantly connected VHN, among which highly perturbed proteins aggregate into an observable core region. The non-core region of VHN forms a large scale but relatively low perturbed periphery. We further validate that the periphery is non-negligible and conducive to identifying comorbidities and detecting drug repurposing candidates for COVID-19. We particularly put forward a flower model for COVID-19, SARS and H1N1 based on their peripheral regions, and the flower model shows more correlations between COVID-19 and other two similar diseases in common functional pathways and candidate drugs. Overall, our periphery-core pattern can not only offer insights into interconnectivity of SARS-CoV-2 VHPs but also facilitate the research on therapeutic drugs.

Key words: COVID-19; SARS-CoV-2; virus-host network; omnigenic; disease module

Introduction

The pandemic of COVID-19 is an acute respiratory infection caused by the 2019 novel coronavirus SARS-CoV-2 [1], and human-to-human transmission of the virus has been confirmed [2]. As of 20 January 2021, more than 96 million COVID-19 cases have been confirmed globally with more than 2 million deaths. But to people's great disappointment, there are still no effective medications for COVID-19. As a consequence of it, COVID-19 has posed an unprecedented threat to all people around the world. Therefore, the pressing situation entails systematic understanding of the molecular interaction mechanism of the disease, so as to provide scientific strategies for developing effective vaccines and therapeutic drugs.

In a molecular interactions network (human interactome), nodes are genes and edges are built on observational inference of

Bingbo Wang received his PhD from Xidian University in 2014. He is an Associate Professor of the School of Computer Science and Technology, Xidian University. His research interests are in bioinformatics and network medicine.
Xianan Dong conducted analysis of disease modules in biological networks, and developing statistical methods and bioinformatics tools. She is an Assistant Researcher of the School of Computer Science and Technology, Xidian University.
Jie Hu's research interests are in the area of network medicine and computer programs. She is an Assistant Researcher of the School of Computer Science and Technology, Xidian University.
Xiujuan Ma is a principle software developer in the School of Computer Science and Technology, Xidian University.
Chao Han is a graduate student of the School of Computer Science and Technology, Xidian University, and her research focus is bioinformatics.
Yajun Wang is a lecturer of the School of Humanities and Foreign Languages, Xi'an University of Technology. His research focus is on data science.
Lin Gao received her PhD from Xidian University in 2003. She is a Professor of the School of Computer Science and Technology, Xidian University. Her research focus is bioinformatics.
Submitted: 27 November 2020; **Received (in revised form):** 30 March 2021

their interactions responsible for specific cellular functions [3]. A disease perturbed gene represents a node, whose perturbation (mutations, deletions, copy number variations or expression changes) can be linked to a particular disease phenotype [4]. Accordingly, the degree of its perturbation can be quantified by frequency or fold change indexes. A disease is caused by the interplay of multiple molecular processes rather than by an abnormality in a single gene. Hence, it is of great significance to study molecular mechanism of the disease in the context of human interactome, a comprehensive map of all biologically relevant molecular interactions. According to the disease module hypothesis [4], a disease represents a local perturbation of the underlying disease-associated subgraph. Such perturbations could represent the removal of more proteins (e.g. by nonsense mutations), the disruption or modifications in the strength of their interactions, producing recognizable developmental and/or physiological abnormalities. Driven by high-throughput interactome mapping efforts [3] and the wealth of genome-wide genetic association data [5], the emerging tools of network medicine provide a platform for systematically exploring the molecular complexity for specific diseases [4, 6–9]. The aggregation of disease proteins has been supported by a range of biological and empirical evidence [10, 11] and has promoted the development of many tools to identify disease modules [12–14]. Disease module plays a vital role in uncovering the molecular mechanism of disease causation, identifying new disease comorbidity [15] and aiding rational drug target identification [16, 17].

In addition, some scholars have recently advanced an entirely new view of diseases from polygenic to omnigenic [18, 19]. Boyle *et al*. [18] observed that disease-causing variants do not cluster into local subnetworks (mesoscale disease module) that drive disease etiology, but association signals tend to be spread across most of genes (macroscale disease neighborhood). The contribution of genes to diseases can be divided into the direct role of core genes and the indirect role of peripheral genes [20]. Boyle *et al*. proposed that gene regulatory networks are so sufficiently interconnected that all genes expressed in disease-relevant cells are liable to affect the functions of disease-related core genes. And the most heritability can be explained by effects on genes outside core pathways. They referred to this hypothesis as an 'omnigenic' model [18]. The key questions and tests suggested by this model are: How should core genes be defined? How many distinct periphery genes would contribute to core genes? Can we infer the effects of peripheral genes from their relation to core genes? Recently, some instructive methods have been developed to define and identify core genes based on genetic and topological properties [19, 21–23].

Wang *et al*. [24] tried to infer the effects of peripheral genes from their relation to core genes and introduced a seminal framework, which can detect peripheral and core regions of a disease based on the local maxima of connectivity significance between the deferentially expressed genes in the human interactome. They have provided evidence that core genes are more enriched for Genome-wide association study (GWAS) data [5] and Online Mendelian Inheritance in Man (OMIM) data [25], while periphery shows relationship between diseases through their overlapped regions. Core region typically consists of genes specific for the underlying disease. Even for a pair of similar diseases, the size of overlap between their cores is not significantly large (compared 1000 randomized subnetworks). They attempted to make an assumption that the similar molecular mechanism of similar diseases lies in their common peripheries. The overlapped periphery is helpful in identifying the molecular mechanism of disease causation, new comorbidity and aiding rational drug target. They have proposed a novel flower model to explain the organization of genes, with (specific) petals representing core of different diseases and the (shared) stem representing the periphery. Overall, the framework is able to demonstrate the hypothesis proposed by Boyle *et al*. [18] and might help address numerous problems with respect to disease gene identification, drug repositioning, and provide an insight into a general understanding of human complex diseases.

As for COVID-19, there have been some studies on drug repurposing based on these network medicine platforms. Zhou *et al*. [26] presented an integrative, antiviral drug repurposing methodology. Employing a systematic pharmacology-based network medicine platform, they quantified the interplay between the virus-host proteins (VHPs) and drug targets in the human protein–protein interaction network [26]. Gysi *et al*. [27] used 332 COVID-19 proteins from David *et al*. [28] as disease module. They predicted the drug candidates for the treatment by defining the location of the disease module within the human interactome [27]. They summarized the result of strategies based on network proximity, network diffusion and artificial intelligence, thereby to arrive at 81 promising repurposing candidates. But there are still several problems to be solved. The network proximity and diffusion methods offer low accuracy of drug prediction (AUC scores about 0.6), which needs to be further improved. They also found COVID-19 proteins that do not overlap with disease proteins associated with any major diseases. This makes it difficult to measure accurately the relationship between COVID-19 and other diseases. These results were most likely due to the fact that they only considered 332 highly perturbed proteins (28.6%), whereas there were 1160 human proteins (Saint_BFDR $\leq$ 0.05, Saint_BFDR is Bayesian False Discovery Rate reported by SAINTexpress [29]) affected by coronavirus SARS-CoV-2. A large amount of information is ignored, resulting in inadequate prediction of COVID-19 comorbidity and drugs. Therefore, from the perspective of omnigenic, to explore peripheral and core regions that govern the underlying perturbation mechanism of COVID-19 is desirable and valuable: (i) can systematically study the overall perturbation of the virus to the cell based on more comprehensive information and (ii) as well as probe the new pattern to improve current comorbidity analysis and the predictive effect of drug repurposing.

In this paper, we study human proteins which interact with SARS-CoV-2 viral genes during infection [28] to disclose the periphery and core regions of COVID-19. Firstly, we identify 934 peripheral proteins and 78 core proteins based on topological connectivity. The combination of peripheral proteins and core proteins and their interactions in the human interactome form the Virus-Host Network (VHN) of COVID-19. And we find these proteins all have high centrality and VHN and its core region have significantly high cohesiveness. Furthermore, VHN and core regions are used to analyze the relationship between COVID-19 and other diseases. Strong correlations between COVID-19 and other diseases are found, including SARS, H1N1 and cancers, as well as immune system and nervous system diseases. Subsequently, taking peripheral regions into consideration, we use a network-based framework to predict drug targets and offer highest AUC: 0.77. Finally, we use flower model to show relationships between COVID-19, SARS and H1N1, and significant related disease genes, drug targets and functional pathways are detected in their overlapped proteins. The VHN and core region of COVID-19 help analyze statistically relevant diseases and improve the accuracy of drug prediction. In a nutshell, adopting the tools of network science in studying COVID-19 would provide us with new insights into disease relationship and novel therapies and drugs.

Results

Peripheral and core regions of VHN

SARS-CoV-2 infects human cells by hijacking the host's translation mechanisms to generate viral proteins, which bind with multiple human proteins to initiate the molecular processes required for viral replication and additional host infection [30]. The human interactome (see Materials and Methods, Table S1 available online at https://github.com/wangbingbo2019/ENCORE-of-COVID-19) includes a variety of known physical interactions in human cells. The proteins which interact with SARS-CoV-2 viral genes during infection [28] are called VHPs, and we map these proteins onto the human interactome. From the perspective of omnigenic module, we identify the peripheral and core regions of COVID-19 and emphasize the important role of the periphery in analyzing disease relationships and drug repurposing. We select the VHPs with Saint_BFDR $\leq$ 0.05 in SAINTexpress [29] (see Materials and Methods), obtain 1160 VHPs and compute the Largest Connected Component (LCC) induced by the VHPs. We distinguish between the peripheral and core regions of COVID-19 based on the local maxima of connectivity significance between the VHPs (see Materials and Methods). MIST score [31, 32] is used to measure the degree of SARS-CoV-2 disturbance to a protein. At different MIST score cutoffs (see Materials and Methods), we select the corresponding subsets of VHPs and identify the induced LCC, which determines the disease neighborhood for the subset of VHPs. And we compare the size of the LCC with the same number of random proteins in the human interactome and compute the LCC's z-score. We identify two peaks on the least squares polynomial fitting curve of the z-score values representing local maxima of VHPs aggregation (see Materials and Methods, Figure 1A). When MIST score $\geq$ 0.088, we observe a LCC with size being 1012 and LCC's z-score is 14.87 (first maxima), as the VHN. We visualize VHN with Cytoscape [33] (Figure 1B). And, when increasing to a high cutoff MIST score $\geq$ 0.876, we observe a highly perturbed LCC with size being 78 and LCC's z-score is 10.66 (second maxima), as the core region (Figure 1D). Non-core region of VHN (with size 1012–78 = 934 proteins) is considered as peripheral region. VHN contains two detectable inner regions, which are the core and peripheral regions. Both regions have significant connectivity: the core region is characterized by high perturbation and the peripheral region is characterized by large-scale.

David *et al.* discovered 332 high confidence proteins (HC_VHPs) interacting with SARS-CoV-2 viral proteins [28]. Among these, they identified 66 druggable human proteins or host factors targeted by 69 existing FDA-approved drugs and evaluated for efficacy in live SARS-CoV-2 infection assays. We find 98.7% (77/78) core proteins belong to the HC_VHPs, showing that almost all of the core proteins are HC_VHPs with which the viral proteins bind. In addition, in order to further illustrate the biological significance of 78 core proteins, we use the following tests to verification: (i) Drug target enrichment analysis, we collect 62 of the 66 drug targets found in HC_VHPs by David *et al.* [28]; (ii) Enrichment analysis of Differential Expression Genes (DEGs), we obtain 1226 DEGs (Benjamini–Hochberg adjusted *P*-value $\leq$ 0.05) from the work of Mike *et al.* [34]; (iii) Enrichment analysis of related diseases, we summarize the DEGs of three diseases thought to have relation with COVID-19, including lung cancer [35, 36], cardiomyopathy [37–39] and Parkinson's disease [40–42] and (iv) Literature verification. Totally, we find additional evidence that 65.4% (51/78) proteins are related to COVID-19 (Table 1, more details in Table S2 available online at https://github.com/wangbingbo2019/ENCORE-of-COVID-19). And we observe a LCC (156 connected proteins with z-score 5.85) when MIST score $\geq$ 0.699; compared with LCC (165 connected proteins with z-score 5.73) of HC_VHPs, the Jaccard coefficient is 0.91 (Figure 1C). This means that the LCC of HC_VHPs has significantly high connectivity, but it is not the local maxima of LCC's z-score curve; instead, it corresponds to an obvious local minima. Therefore, we believe that the core region with the most significant connectivity should not be defined by the HC_VHPs. For the peripheral region with large scale, it not only contains more virus host proteins but also forms a significantly connected network area with non-negligible information. Next, we investigate whether different regions have different topological characteristics or play quite different roles in comorbidity and drug repurposing.

Topological characteristics of VHN and core regions

Here, to better understand the interaction patterns of core and peripheral regions, we use human interactome (see Materials and Methods) as a background network to analyze the basic topological properties of the VHPs. We factored into Degree, betweenness [43, 44], closeness [44] and clustering coefficient [45]. Compared with other proteins in the network, both 78 core proteins and 1012 VHN proteins have a significantly high centrality (Wilcoxon rank-sum test *P*-value $<$1.0e$-$16, Figure 2A). This indicates that VHPs tend to be the hub nodes in the network; viruses tend to influence the network hubs to expand their influence on the whole system. The only exception is that the clustering coefficient of 78 core proteins is not significantly high, showing that the 78 core proteins do not tend to participate in the locally triple structure (Figure 2A). In short, viruses have a tendency to infect the hub proteins in the network to quickly affect the entire system.

In addition, since VHPs have a significantly high centrality (Figure 2A), we wonder whether they tend to form inwardly compact module. Respectively, we compare internal and external connectivity indexes (see Materials and Methods) of 1012 VHN proteins and 78 core proteins with 1000 randomly connected components (as randomized counterparts, see Materials and Methods), and calculate the significance based on z-score. For VHN (Figure 2B), internal connectivity is significantly high (z-scores $>$25), and the significance of external connectivity is correspondingly low (z-scores are about 9). For core, internal connectivity is slightly high (z-scores are about 1), while external connectivity is slightly low (z-scores are about $-$0.9, Figure 2B), suggesting that the internal and external connectivity of core are broadly in line with randomized counterparts. Then, we combine internal and external connectivity for in-depth analysis. Conductance or cohesiveness accordingly measure the fraction of total external or internal edge volume of the node set (see Materials and Methods). The entire VHN is highly aggregated with significantly different conductance (z-score <-25) and cohesiveness (z-score $>$ 25) with randomized counterparts. But the absolute value of z-scores of core is reduced from 25 to about 4, indicating that core occupies more external but less internal interactions in VHN. As a conclusion, we get an aggregated VHN, in which signals gather. In the VHN, core extends and tends to receive signals from the peripheral region. Highly perturbed core not only interacts internally, but also tends to interact with the wider peripheral region, validating the necessity of paying attention to external effect.

Comorbidity of COVID-19 based on VHN

Most COVID-19 patients show mild-to-moderate symptoms, a few are asymptomatic, but some patients with underlying

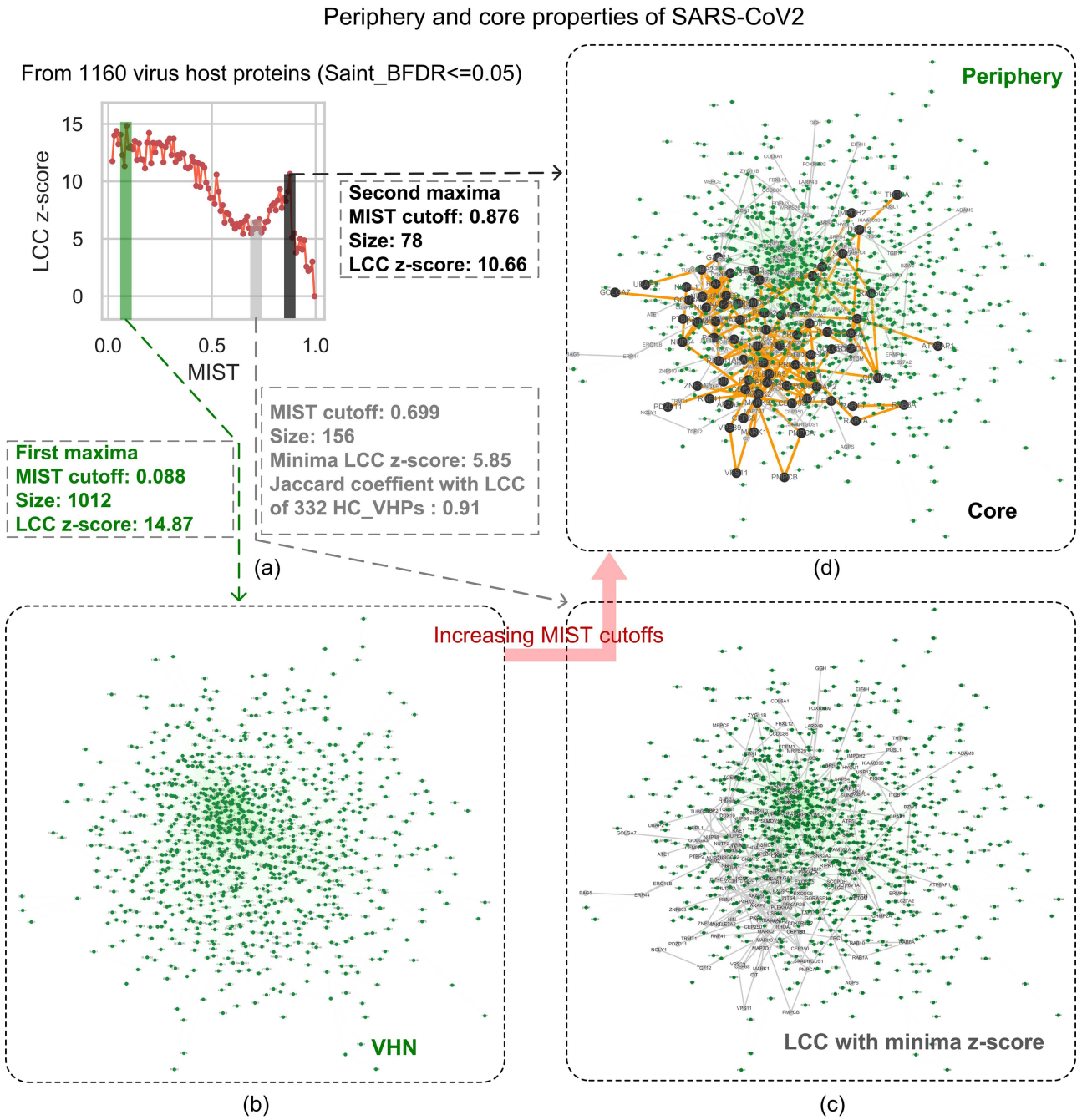

Figure 1. Peripheral and core regions of VHN. (**A**) Detection of peripheral and core regions. MIST score indicates the degree of disturbance of the host proteins by the virus, and the larger the score, the stronger the disturbance. The z-score is calculated from the LCC of the VHPs and the random LCC. Red nodes represent the LCC's z-scores of the VHPs, MIST score of which is greater than threshold. The ever-increasing MIST threshold corresponds to the process of focusing our attention from the whole VHPs to its high disturbance area. This curve shows a bimodal trend, a significant connected component of a large number of proteins appearing at a low disturbance threshold, and then z-scores significantly drop to a trough, and then a significant connected local maximum appears at a high disturbance threshold. The first maximum value indicates VHN (green bar), and the second maximum value indicates a detectable core area (black bar). The gray bar corresponds to local minimum. (**B**) VHN is a connected graph formed by 1012 proteins, with green nodes representing proteins. (**C**) The LCC (156 gray nodes) with local minima LCC's z-score which is highly consistent with the LCC of 332 highly confidence proteins. (**D**) The peripheral proteins (green) and core proteins (black) of VHN. The interactions between core proteins are highlighted by orange.

diseases develop severe pneumonia and even severe comorbidities. Analyzing the relationship between COVID-19 and other diseases will be of great importance in offering insights into the mechanism of the disease. And combination of basic research on the disease with actual diagnostic detection and drug treatment will greatly promote the prevention and treatment of COVID-19. First, we identify the disease modules of 72 well curated diseases (see Materials and Methods). Second, we check the overlap between the disease modules with COVID-19. Then, we define the similarity sim_{AB} between diseases based on the

Table 1. Disease biology of 78 core proteins

Protein	Disease bBiology	Protein	Disease biology	Protein	Disease biology
VPS11	□	PPIL3		NUP214	★□
ATP6AP1	★□	G3BP2		AP3B1	
UBAP2	□	MIB1	△□	ARF6	
IMPDH2	★	PKP2	□	MIPOL1	
RIPK1	★□	RAB7A	□	SCCPDH	
MYCBP2	○□	PLEKHA5	□	AKAP9	□
MARK1	△	RAB1A	□	G3BP1	○
MARK3	★□	RAB2A	○△□	GOLGA7	□
CWC27		GOLGA2	□	CEP250	★□
ERC1	□	GOLGA3	□	PRIM1	
PDZD11		GOLGB1	□	PRIM2	
PSMD8		RAB5C		PDE4DIP	□
NUP62	★	STOML2	□	PRKACA	★
ZNF318	□	P4HA2		CEP135	
NUPL1	★△	USP13	□	NINL	
POLA2	□	RAE1	★□	PRKAR2A	
CHMP2A	□	TAPT1		PRKAR2B	△□
RAB8A	□	PMPCB	□	INTS4	
NUP54	★△	VPS39	□	CDK5RAP2	
CEP68		PTBP2	△□	RNF41	
RAB10	△	POLA1	□	MARK2	★□
AP2M1	□	NUP88	★□	USP54	□
AP2A2	□	NUP98	★△□	NUTF2	
PMPCA		TRMT1	□	RBM41	
SUN2	△□	SLU7		PCNT	□
GORASP1	★□	THTPA	△□	NIN	○

★: Drug target; ○: DEGs of COVID-19; △: Genes for diseases associated with COVID-19; □: Literature verification.

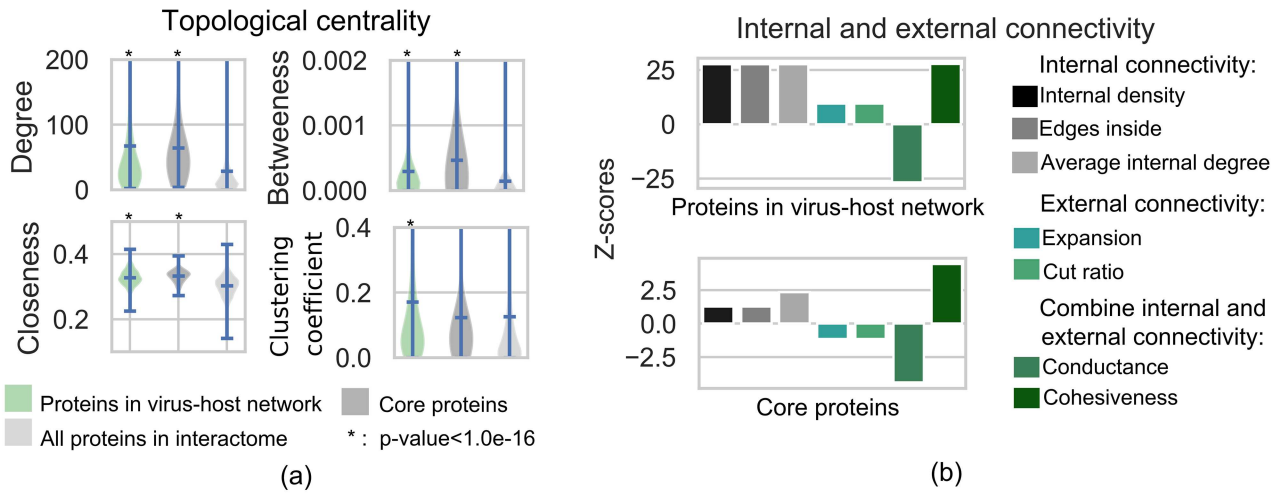

Figure 2. Topological characteristics of VHN and core region. (**A**) The comparison of degree, betweenness, closeness and clustering coefficient of the VHN proteins (green), core proteins (dark gray) and all proteins in human interactome (light gray). *P*-value is given by the Wilcoxon rank-sum test, which is used to quantify the topological differences between proteins in the core or peripheral regions and all proteins in the human interactome, where * in the figure represents *P*-value <1.0e−16. (**B**) The internal and external connectivity of VHN and core. Measures of internal and external or combine connectivity are shown in the legend.

distance of their location in the network (see Materials and Methods). The empirical *P*-value of the size of overlap and similarity are given by conducting 100 random experiments. Finally, we sort diseases according to their similarities to COVID-19 and get the ranking list of comorbidities.

We use two strategies to collect disease modules. On the one hand, we obtain 70 disease modules constructed in OMIM and GWAS studies [5, 12, 25]. These disease modules do not form connected subgraph in the human interactome, so we employ C3 algorithm [13] (see Materials and Methods) to get 70 connected disease modules. In addition to these 70 diseases, we are also interested in the relationship between COVID-19 and two highly infectious respiratory diseases, SARS and H1N1. We get few proteins associated with SARS (only one disease protein) and H1N1 (no disease protein) from OMIM and GWAS studies. The number of proteins is too small to be used as

disease modules to analyze disease relationships. Therefore, on the other hand, we use gene expression profiling of patients with SARS [46] and peripheral blood cells expression data form H1N1 influenza patients [47]. We conduct differential expression analysis by GEO2R tool [48] with healthy controls and obtain the corresponding DEGs. Then, we compute LCC's *z*-scores of increasing fold change cutoffs and detect the peripheral and core regions for SARS and H1N1 (see Materials and Methods). The core regions of SARS and H1N1 are used as disease module for subsequent disease relationship analysis. In total, we collect 72 connected disease modules (details in Table S2 available online at https://github.com/wangbingbo2019/ENCORE-of-COVID-19) with 2913 disease proteins, including SARS and H1N1 (see Materials and Methods). These disease modules are utilized to analyze the disease relationship with COVID-19.

We calculate the size of overlap between 72 disease modules with the VHN and core of COVID-19. Based on core region of COVID-19, we find only 16 of 72 diseases have overlapped proteins, and the number of overlapped proteins is a maximum of 3 (details in Table S3 available online at https://github.com/wangbingbo2019/ENCORE-of-COVID-19). Most disease modules (77.8%) have no overlapped proteins with core of COVID-19, and compared with 100 tests between randomly connected components, *z*-scores are generally less than 0 (Figure 3A). This indicates that there is no trend of sharing disease proteins between the core of COVID-19 and other disease modules. However, there are seven disease modules significantly sharing proteins (*z*-score > 1) with COVID-19's core region (details in Table S3 available online at https://github.com/wangbingbo2019/ENCORE-of-COVID-19), including cardiac arrhythmia, myeloproliferative disorders, cardiomyopathies, blood platelet disorders, basal ganglia diseases, liver cirrhosis and Myeloid leukemia. Among the seven diseases, cardiac arrhythmia shares the most proteins (*RAB5C*, *PRKACA* and *AKAP9*) with COVID-19 (*z*-score = 3.16). Although shared proteins are few in number, we do further analysis of their annotations in GeneCards [49]. For example, diseases associated with *RAB5C* include Argentine hemorrhagic fever, among its related pathways are metabolism of proteins and innate immune system. Gene Ontology (GO) annotations related to this gene include GTP binding and GDP binding. The related diseases, pathways and GO annotations of these shared proteins uncover some molecular mechanism of COVID-19. Considering the role of periphery, we expand the scope from core to the entire VHN to present the overlapped proteins with other disease modules, the number has been greatly improved. At most, there are 57 overlapped proteins with the nutritional and metabolic diseases module, but *z*-scores (see Materials and Methods) for significance are mostly less than 0, even lower than the *z*-scores of the core region (Figure 3A). This suggests that there is no tendency for VHN and disease modules to share disease proteins.

Based on the assumption that similar diseases are closer to each other in the human interactome, we use the location of the disease modules in the network to compute their network-based distances. Based on this, we measure the disease similarities (see Materials and Methods) between 72 disease modules and core of COVID-19. In addition, *z*-scores and *P*-values are obtained by comparing with 100 randomly connected components. We show the rank list of diseases by similarities based on core region (details in Table S3 available online at https://github.com/wangbingbo2019/ENCORE-of-COVID-19). The top seven diseases are Basal ganglia diseases, SARS, Myeloid leukemia, Colorectal neoplasms, Motor neuron disease, Psoriasis and Cardiomyopathies. The result is consistent with existing research: cancer whose comorbidity in COVID-19 patients is well documented [50, 51]. COVID-19 may directly affect the immune system [52] and nervous system [53]. In addition, the epidemiological characteristics of COVID-19 are similar to those of SARS [54].

The analysis of the overlap and distance between core region of COVID-19 and other disease modules provides a lot of information on disease relationship. Furthermore, we expand our vision to the entire VHN, hoping to gain more information about the disease relationship. Since there are no statistically larger (*z*-scores are mostly less than 0) overlap between VHN and other disease modules (details in Table S3 available online at https://github.com/wangbingbo2019/ENCORE-of-COVID-19), we still use the network-based distance between 72 disease modules and VHN as a measure of disease similarity. The closest seven diseases are Nutritional and metabolic diseases, SARS, Crohn disease, Lymphoma, Myeloid leukemia, Type 2 diabetes mellitus and systemic Lupus erythematosus (Figure 3B, Table S3 available online at https://github.com/wangbingbo2019/ENCORE-of-COVID-19). After considering peripheral proteins, the similarity between SARS and COVID-19 is still high ($sim_{AB} = 0.36$), ranking second in the list of diseases (details in Table S3 available online at https://github.com/wangbingbo2019/ENCORE-of-COVID-19). In addition, comparing the top seven diseases on similarity diseases list based on VHN with those based on core, we obtain additional comorbidities of COVID-19, including Crohn disease, Lymphoma, Type 2 diabetes mellitus and Systemic lupus erythematosus. The comorbidities of these four diseases with COVID-19 have been documented [55–58]. Based on VHN to calculate the similarity between COVID-19 and other diseases, additional comorbidities of COVID-19 can be found, indicating that VHN can provide further advantageous information for predicting comorbidities of COVID-19.The locations of the 72 disease modules and VHN and core region of COVID-19 in the human interactome enable us to identify similar diseases, whose molecular mechanisms overlap with SARS-CoV-2 targeted cellular processes, which allows us to predict potential comorbidities with COVID-19.

Drug repurposing for COVID-19

The situation is getting grim as COVID-19 is rampant around the world, and effective drugs and vaccines are in urgent need. However, the traditional drug development process is too long to meet the urgent need for COVID-19 treatment; therefore, it demands the rapid identification of drug-repurposing candidates. The network medicine approach has already offered a promising framework to accelerate drug discovery [59], helping us quantify drug–disease relationships [60–62]. Network-based approaches show that most drugs do not target directly disease proteins, but perturb the proteins within or in the immediate vicinity of the corresponding disease module [61, 63]. In order to predict drugs that can be used in the treatment of COVID-19, we use a network proximity method, which quantifies the relationship between VHPs and drug targets in the human interactome (see Materials and Methods).

We obtain 4428 drugs and 2256 targets from the work carried out by Cheng *et al*. [64] and map these targets onto the human interactome, retaining 4380 drugs and 2161 targets (see Materials and Methods). First, we calculate distance between VHPs and drug targets for analyzing the relationship between the COVID-19 and drugs. Then, in order to eliminate bias, we determine the statistical significance for the observed proximity by *z*-score. We construct a reference distance distribution between a randomly selected group of proteins with matching size and

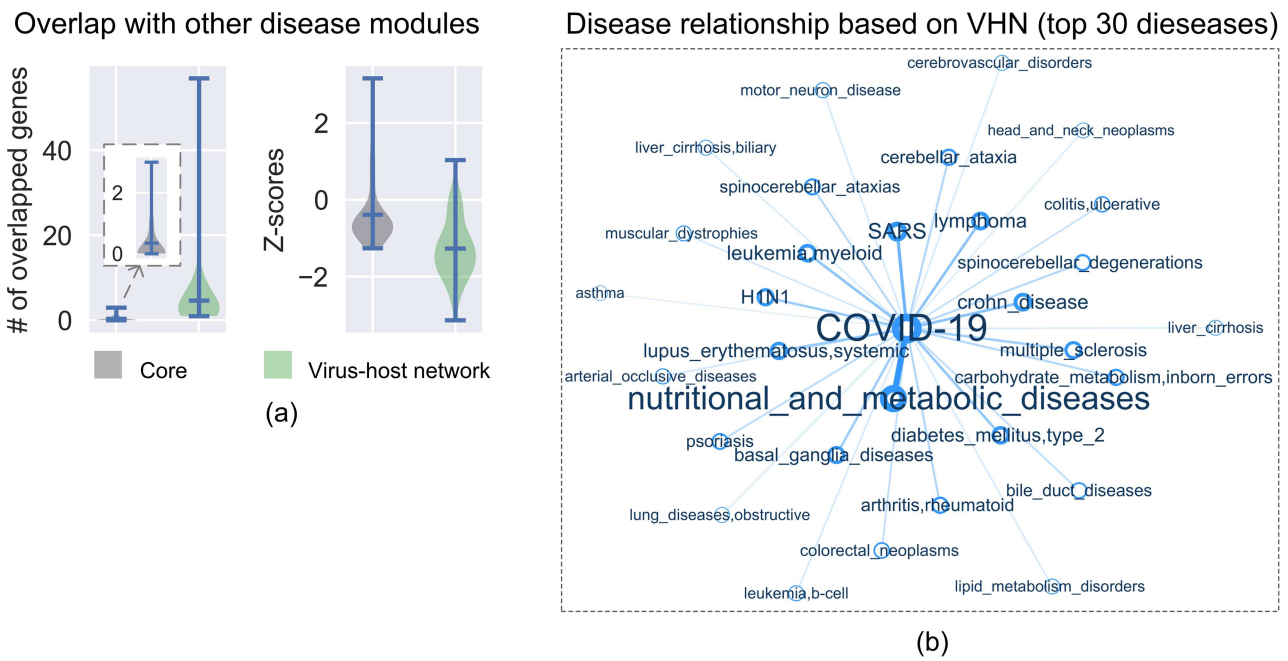

Figure 3. Disease relationship. (**A**) The number of overlapped proteins of core and VHN with other disease modules. Z-scores are given by randomly selecting a set of proteins of the same size 100 times. (**B**) The disease relationship is based on periphery. Only the top 30 diseases are shown. The figure represents each disease as a circle and the disease font size and the thickness of the edge are directly proportional to the similarity values. The smaller the font size and the thinner the edge, the lower the similarity with COVID-19.

degree distribution as the VHPs and drug targets (see Materials and Methods). Z-scores are obtained by comparing the observed proximity to the reference distance distribution. The smaller the z-score, the closer the distance between the corresponding targets of these drugs and VHPs, implying that the retained drugs may be applied to the treatment of COVID-19. In total, we end up with 4380 drug rankings and their z-scores. To evaluate the predictive power, we test its ability to recover the drugs currently in clinical trials for COVID-19 treatment. We use a list of 67 drugs currently undergoing clinical trials from ClinicalTrials.gov (details in Table S4 available online at https://github.com/wangbingbo2019/ENCORE-of-COVID-19) as a gold standard. Using drugs rankings, we calculate TPR (True Positive Rate) and FPR (False Positive Rate) according to different z-score thresholds. Plot the Receiver Operating Characteristics (ROC) curves [65] and calculate the Area Under the Curve (AUC) scores for performance analysis. As Figure 4 shows, as the z-score threshold decreases, the AUC increases, indicating that the prediction accuracy of the drug that meets the criteria is improved. It verifies our hypothesis that a drug may perturb COVID-19 when its targets are located in the VHPs neighborhood. The best AUC score occurs in the experiment based on VHN proteins, the highest accuracy is 0.7 when z-score < -1.5 (Figure 4). The results based on core proteins are not satisfactory (0.47–0.55, Figure 4), which necessitates adding peripheral proteins to drug-repurposing. This performance suggests that taking the peripheral region into consideration can significantly improve the effect of drug-repurposing and provide reusable drug candidates for the prevention and treatment of COVID-19. Finally, combining the AUC score and the number of drugs, we select the top 1000 drugs (details in Table S5 available online at https://github.com/wangbingbo2019/ENCORE-of-COVID-19) in the ranking of drugs based on the distance between VHN and drug targets, representing our final repurposing drugs list for COVID-19.

Flower model for relationship between COVID-19, SARS and H1N1

As COVID-19 spreads globally, the epidemiological features of the disease are being revealed. COVID-19, SARS and H1N1 are all malignant infectious diseases, causing huge losses to human society. Previous analyses of disease relationships based on network distance have shown some similarities between COVID-19, SARS and H1N1. Since we have obtained the peripheral and core regions of these three diseases, we want to further study the disease relationship of COVID-19, SARS and H1N1 in detail by using a flower model (Figure 5A). The disease neighborhood of COVID-19 is composed of peripheral and core regions with sizes of 934 and 78, and the disease neighborhoods of SARS and H1N1 are composed of peripheral and core regions with sizes of 889, 2209 and 60, 93, respectively (Figure 5B). We model their locations in the human interactome as flower model and focus on the 500 proteins, a subset containing the core regions of the three diseases and the common peripheral regions of COVID-19, SARS and H1N1. Specific core region is shown as black, blue and red petals, while overlapped peripheral region is shown as green stem (Figure 5A). This flower model can help us uncover the relationship between COVID-19, SARS and H1N1. The details of similarity of three diseases are shown in Supplementary Materials (detail in Table S6 available online at https://github.com/wangbingbo2019/ENCORE-of-COVID-19).

We map the peripheral and core proteins of COVID-19, SARS and H1N1 onto the human interactome and discover that these three disease neighborhoods have both specific and common regions (Figure 5A). The three specific core regions are independent of each other and have no overlap, while their peripheral regions overlap each other. The sizes of overlap of VHN with SARS and H1N1 are 107 and 235, respectively, and size of overlap between SARS and H1N1 is 216. In addition, the size of overlap between the peripheries of these three diseases is

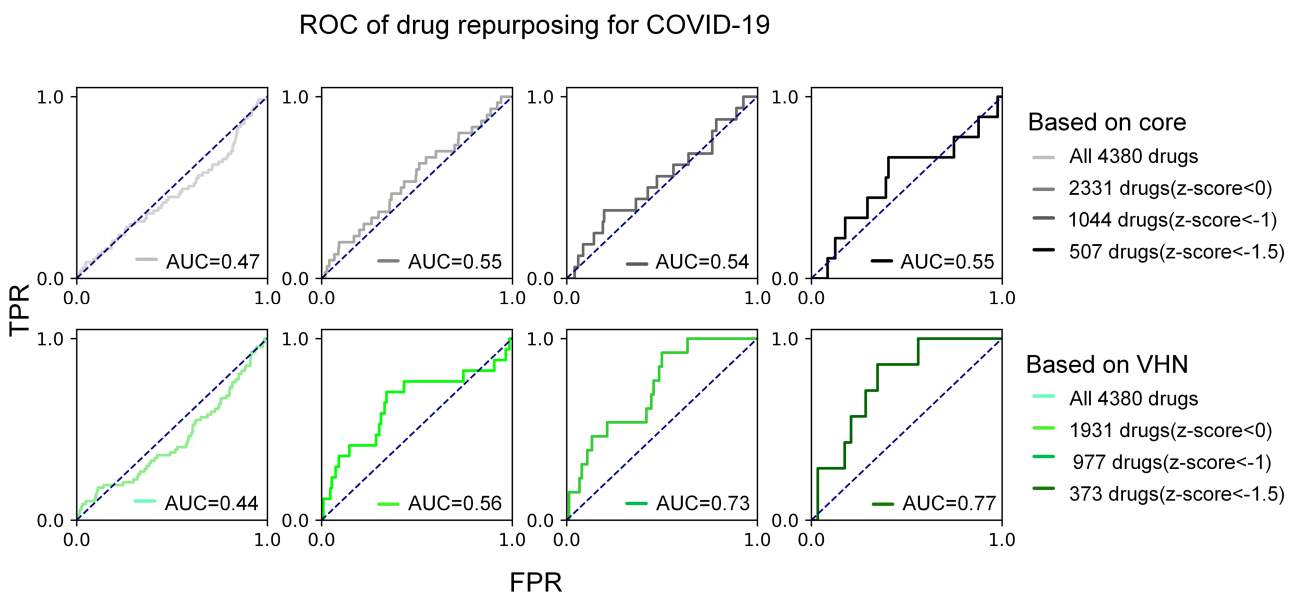

Figure 4. Performance analysis for drug repurposing. ROC Curves and AUC scores based on core region (gray) and VHN (green) for reducing drug z-scores obtained by network proximity strategy. The deeper the color, the smaller the z-scores; it indicates that the drug targets and the disease neighborhood are closer in the network. The dotted line represents the performance of the random classifier and AUC is 0.5.

32. COVID-19, SARS and H1N1 have a statistically significant large common peripheral region, which suggests a possible common molecular mechanism between them. This indicates that taking the peripheral region of disease into consideration significantly improves the prediction of comorbidities among complex diseases. In addition to the overlapped proteins between diseases, we want to explore whether the peripheral region contains other potential information, which may affect the relationship between diseases. Therefore, a topological measure is used to detect the mediator proteins, which mediate the molecular interactions between diseases (see Materials and Methods, Table S6 available online at https://github.com/wangbingbo2019/ENCORE-of-COVID-19). Mediator proteins may not be part of either disease module, but they are located in the shortest paths connecting the two diseases module, playing a key role in mediating the interaction between the two diseases [14]. We find key nodes that mediate diseases interactions in the peripheral regions of these three diseases, which further shows that considering the peripheral regions of the diseases is more effective in analyzing the relationship between the diseases (Figure 5A).

Overlapped peripheral proteins and mediator proteins in these disease neighborhoods may provide us with critical information on disease relationships, thus providing an opportunity to understand the shared molecular mechanisms of these diseases. In order to validate the biological relevance of these proteins, we conduct KEGG pathway enrichment analysis through ConsensusPathDB (CPDB) [66] for Overlapped Periphery of COVID-19 and SARS (OP_C&S), Overlapped Periphery of COVID-19 and H1N1 (OP_C&H), Mediator proteins between COVID-19 and SARS (M_C&S) and Mediator proteins between COVID-19 and H1N1 (M_C&H). As a result, we present 22 KEGG pathways (q-value < 0.01) (Figure 5C). Four proteins sets are consistently enriched in four diseases: Parkinson Disease, Huntington Disease, Alzheimer Disease and Non-Alcoholic Fatty Liver Disease pathways. Of them, Parkinson Disease, Huntington Disease and Alzheimer Disease are neurodegenerative diseases caused by the death of neurons in the brain, leading to cognitive and behavioral dysfunction; they share many similarities at the cellular and molecular levels and key characteristics [67]. Both COVID-19 and SARS are acute respiratory infection caused by coronavirus. Previous studies have shown that although coronavirus is recognized primarily as a respiratory pathogen in humans, its affinity with the basal ganglia suggests its possible role in human Parkinson Disease. There may be an association between coronavirus and Parkinson Disease or other neurological diseases [68]. Furthermore, peripheral inflammation caused by COVID-19 may have a long-term impact on the recovery from the disease, leading to chronic medical conditions such as neurodegenerative diseases [69]. H1N1 influenza virus infections may be associated with central nervous system pathology. Central nervous system inflammation has been implicated in neurodegenerative diseases including Parkinson Disease [70]. And neuronal cells can be infected by pandemic H1N1 viruses [71]. In addition, there is growing evidence that patients with COVID-19 often develop liver damage [72]. COVID-19 is frequently associated with different degrees of abnormal liver function tests and patients with Non-Alcoholic Fatty Liver Disease may have a higher risk of developing severe COVID-19 [73]. The functional enrichment analysis results of overlapped and mediated proteins provide us with a lot of information, which is critical to uncovering the molecular mechanisms of disease and designing effective treatments, and it provides an opportunity to understand the relationship between COVID-19 and other diseases. Then, by analyzing the proteins that are enriched in Parkinson Disease, Huntington Disease, Alzheimer Disease and Non-Alcoholic Fatty Liver Disease pathways, we obtain 58 significantly enriched proteins, functionally enriched P-values are almost all less than 0.025. And by showing in detail the subnetwork of 58 proteins enriched into diseases in the human interactome, we find two distinct complexes, ATP5 and NDUF (Figure 5D). Furthermore, a phenomenon is that the ATP5 complex contains the common peripheral proteins of the three diseases, while NDUF mainly includes the common peripheral proteins of COVID-19 and H1N1.

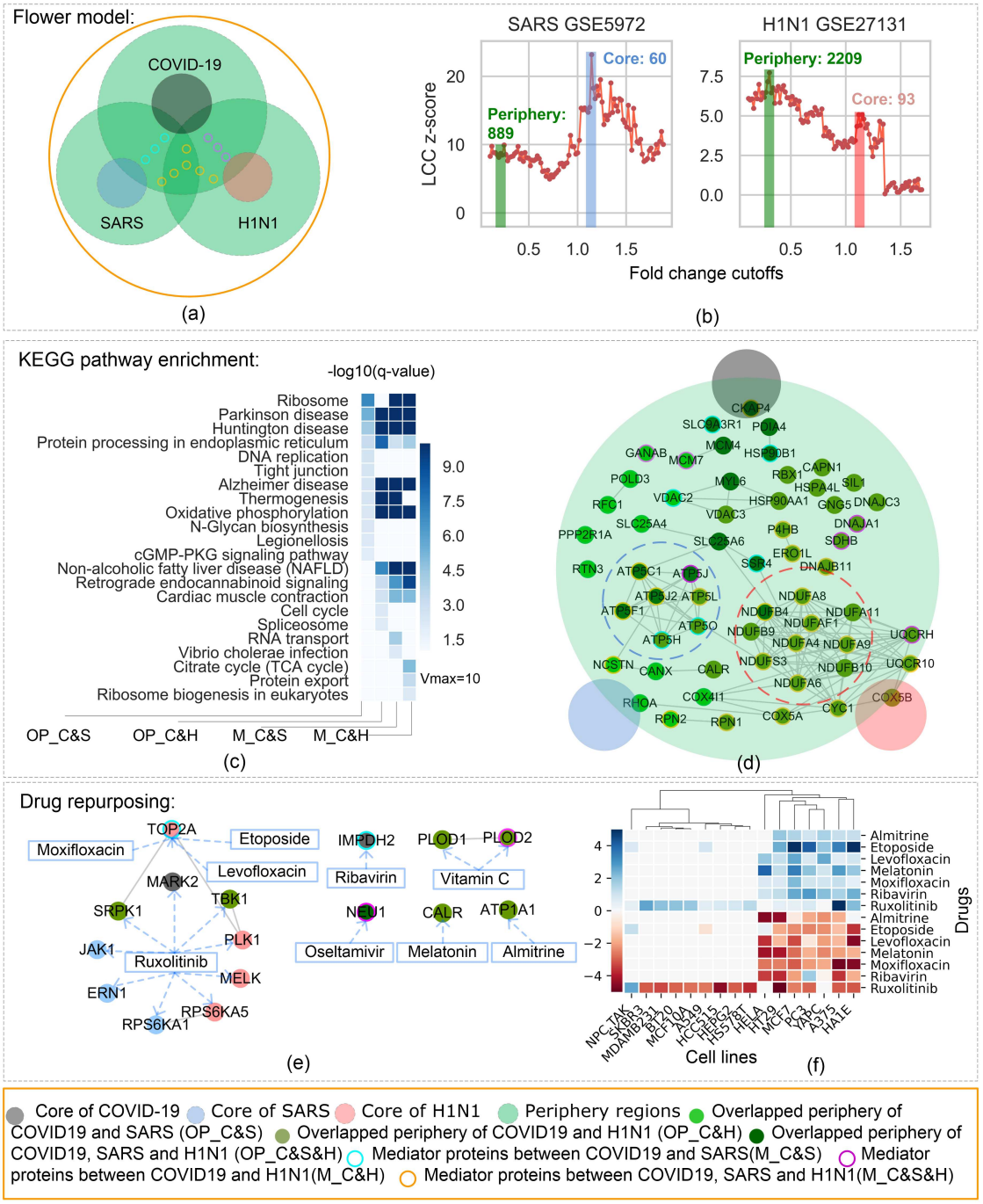

Figure 5. Flower model for COVID-19, SARS and H1N1. (**A**) Schematic diagram of flower model for COVID-19, SARS and H1N1. Black, blue and red represent the specific core region of COVID-19, SARS and H1N1, and green represents their peripheral regions. Light green represents the non-overlapped regions, dark green represents the overlapped regions of three diseases and two mild greens represent the overlapped regions where COVID-19 overlaps with SARS and H1N1, respectively. The darker the green is, the more diseases overlap peripheral region. The mediator proteins of COVID-19 and SARS are circled in blue, the mediator proteins of COVID-19 and H1N1 are circled in purple, and the common mediator proteins among the three diseases are circled in gold. (**B**) The LCC's z-scores with increasing fold change cutoff as threshold of SARS (GSE5972, left) and H1N1 (GSE27131, right) DEGs; then, we detect the peripheral and core regions of SARS and H1N1, respectively. (left) For SARS, we detect LCCs of size 949 and 60, with z-score of 9.73 and 23.17, and thresholds of 0.39 and 1.14. (Right) For H1HI, we detect LCCs of size 2302 and 93, with z-score of 7.74 and 5.09, and thresholds of 0.31 and 1.15. (c) The KEGG pathway enrichment of OP_C&S, OP_C&H, M_C&S and M_C&H. The shades of blue indicate the −log10 operation value for the *q*-value of pathways obtained from the enrichment. In order to show the best effect, set the maximum color depth $V_{max} = 10$ (details in Table S6 available online at https://github.com/wangbingbo2019/ENCORE-of-COVID-19). (**D**) A detailed subgraph of 58 proteins function enriched to diseases and their interactions. (**E**) A detailed diagram of 16 target proteins (dot, the color indicates the region it belongs to) of the nine clinical drugs (rectangular box), the targeting relationship between them is shown by the blue arrow and gray edges are protein interactions. (**F**) Drug response of 7 drugs in 16 cell lines. The top 7 rows: Results of relationship between the VHN proteins and the perturbed proteins caused by the drug. The value in the figure is –log10(*P*-value), where *P*-value is calculated by Fisher's exact test, set the maximum color depth $V_{max} = 5$. The last 7 rows: Results of the spearman correlation ρ between the perturbed proteins caused by the drug and perturbed proteins caused by virus. To better illustrate the results, the value in the figure is $10^{*}\rho$, set the minimum color depth $V_{min} = -5$. The deeper blue in the top 7 rows indicates significantly larger size of overlap, and the deeper red in the last 7 rows indicates stronger negative correlation. We use the Clustermap method of Python package Seaborn [95] for hierarchical clustering, and the specific method used is 'average'.

Among the 500 proteins in flower model, we identify 16 proteins targeted by nine existing drugs currently undergoing clinical trials from ClinicalTrials.gov [71] (Figure 5E, Table S4 available online at https://github.com/wangbingbo2019/ENCORE-of-COVID-19). These drugs can be used to treat respiratory illnesses (Moxifloxacin [74], Almitrine [75]), influenza (Oseltamivir [76]), viruses (Ribavirin [77]) and bone marrow fibrosis (Ruxolitinib [78]). Levofloxacin is also antibacterial against mycoplasma pneumonia and chlamydia pneumonia [79]. Etoposide is a commonly used drug for small cell lung cancer [80]. Thus, flower model can not only reveal the relationship between diseases but also effectively identify potential drugs based on common periphery.

Validation of drug effectiveness

We generate the repurposing drugs list based on the distance between the drug targets and VHN (details in Table S5 available online at https://github.com/wangbingbo2019/ENCORE-of-COVID-19) in the human interactome. This list includes 11 clinical drugs that were not predicted by Gysi *et al.* [27]. And we identify 16 proteins targeted by 9 clinical drugs in the 500 proteins contained in the flower model (Figure 5E). There are 4 duplicate drugs between these 9 drugs and 11 drugs (details in Table S6 available online at https://github.com/wangbingbo2019/ENCORE-of-COVID-19). In order to certify the effectiveness of the peripheral and core regions and the flower model for identifying comorbidities and drug repurposing candidates of COVID-19, we use expression data to illustrate the effect of these drugs on the peripheral proteins and core proteins of COVID-19.

In order to verify the effectiveness of a drug in treating diseases, it is necessary to test whether the drugs can produce the correct perturbation in the cell. We retrieve gene expression perturbation profiles from the Connectivity Map (CMap) database [81, 82], altogether including 861 experimental instances (different drugs, cell lines, doses and time of treatment). In order to measure the effect of each drug on the activity of proteins in the COVID-19 disease module, we measure the size of overlap between the protein products of perturbed genes caused by drug and VHN or core proteins of COVID-19. For example, for sirolimus [83], a potent immune-suppressant, we find that 102 proteins (10%) of VHN have a significantly large size of overlap with protein products of highly perturbed genes (1.0μM) in the lung cell line A549 (Fisher's exact test, FDR-BH $p_{adj} < 0.05$, details in Table S6 available online at https://github.com/wangbingbo2019/ENCORE-of-COVID-19). What is more, we find that there are 120 proteins (12%) of VHN having a significantly large size of overlap with perturbations caused by etoposide [80] (3.33 μM), in the cell line A375 (FDR-BH $p_{adj} < 0.05$), the drug is a semisynthetic derivative of podophyllotoxin that exhibits antitumor activity. At the same time, we find that 23 proteins (29.5%) of core region have a significantly large size of overlap with protein products of highly perturbed genes by etoposide [80] (10.0μM) in cell line A375 (Fisher's exact test, FDR-BH $p_{adj} = 3.06E-28$), and observed that 9 proteins (11.5%) of core region have a significantly large size of overlap with perturbations caused by levofloxacin [79] (0.04 μM) in cell line MCF7 (Fisher's exact test, FDR-BH $p_{adj} = 7.79E-07$), the drug is a fluoroquinolone antibiotic, which helps improve activity against gram-positive bacteria commonly implicated in respiratory infection. These results provide us with direct experimental evidence that the drugs repurposing candidates selected by our periphery-core pattern provide novel insights for the treatment of COVID-19.

Next, to further illustrate the validity of the drug prediction results, we verify whether these drugs can counteract the gene expression perturbations caused by the virus SARS-CoV-2. We carry out the same experiment as that of Gysi *et al.* [27], using the 120 DEGs in the SARS-CoV-2 infected samples of the A549 cell line [84]. We use the protein products of DEGs to measure the Spearman correlation ρ between the perturbations caused by the drug and perturbations caused by virus in the A549 cell line, where $\rho < 0$ indicates that the drug could counteract the effects of the virus infection. For example, ruxolitinib [78] is a janus-associated kinase inhibitor used to treat bone marrow cancer, especially intermediate or high-risk myelofibrosis, whose treatment of the lung cell line HCC515 (0.12μM) counteracts the effects of the SARS-CoV-2 infection, resulting in an inverted expression profile (Spearman correlation $\rho = -0.45$, detail in Table S6 available online at https://github.com/wangbingbo2019/ENCORE-of-COVID-19), and it also has a strong negative correlation in the lung cell line A549 (Spearman $\rho = -0.27$, 0.5 μM, 6 h). Apart from it, moxifloxacin [74] (Spearman $\rho = -0.48$, A375, 0.37 μM, 24 h), almitrine [75] (Spearman $\rho = -0.46$, HELA, 10.0μM, 24 h) and levofloxacin [79] (Spearman $\rho = -0.46$, HA1E, 0.12μM, 24 h), these three drugs were not predicted by Gysi *et al.* [27]. This result shows that the additional drugs predicted by VHN have a certain effectiveness in the treatment of COVID-19, thus indicating that the periphery-core pattern is an effective model for analyzing COVID-19.

Furthermore, we analyze the nine clinical drugs detected by the flower module (Figure 5E). We obtain the gene expression perturbation profiles of seven of the nine drugs in the CMap database. In A375, HELA, HT29, MCF7, PC3, YAPC and HA1E7 cell lines (7/16, 43.7%, Figure 5F), drug validation results have two notable features: (i) The perturbed proteins caused by drugs have a significantly large size of overlap with VHN proteins (P-value < 0.05), indicating that these drugs tend to function in the VHN region in the network. (ii) The perturbations caused by drug and virus have negative correlation, showing the inhibitory effect of drug on viruses. In flower model, Ruxolitinib simultaneously targeted the three disease areas of COVID-19, SARS and H1N1 (Figure 5E), including one core protein (*MARK2*) of COVID-19 and three core proteins (*JAK1, ERN1, RPS6KA1*) of SARS, the three core proteins (*PLK1, MELK, RPS6KA5*) of H1N1 and the two proteins (*SRPK1, TBK1*) on their common periphery. This result demonstrates that Ruxolitinib has a strong ability to control these diseases, which is also reflected in the drug response experiment. Ruxolitinib has a significantly large size of overlap with VHN in 43.7% (7/16) cell lines (Figure 5F). At the same time, the perturbation of virus is inhibited in 81% (13/16) cell lines, and the perturbation caused by Ruxolitinib is negatively correlated with the perturbation caused by virus. In comparison, the other six drugs except Ruxolitinib, all targeted few proteins (≤2) in the flower model (Figure 5E). We also observe that these drugs have no obvious effect on VHN on the other nine cell lines. The tissue information of eight cells is known, including five mammary breast cell lines, one liver cell line and two lung cell lines. The effective drug on the liver cell line (HEPG2) is Ruxolitinib, and the effective drugs on the lung cell line (A549, HCC515) are Ruxolitinib and Etoposide. In conclusion, these results provide evidence for the efficacy of the drugs we have predicted for COVID-19, verify the feasibility of the method of peripheral and core regions detection and the flower model based on omnigenic theory to analyze disease relationship and detect drug repurposing candidates for COVID-19, and further demonstrate that considering peripheral proteins could provide a better platform for the study of COVID-19.

Discussion

In this study, we draw upon the latest advances in COVID-19 virus-host research and network medicine methods to identify the VHN and core region of COVID-19, and we find that both VHN and core regions are internally tightly connected topologies in the human interactome. Then, we combine the C3 disease modules of 70 diseases and the core regions of SARS and H1N1 to analyze the disease similarity. We compute their network distance with the VHN or core region of COVID-19 based on their location in human interactome and find several high similarity diseases including immune and neurological diseases and cancers. We identified drug targets based on network proximity and predicted drug outcomes as high as 0.77, suggesting that COVID-19's peripheral and core regions also provide an opportunity for drug repurposing. This result can provide new insights into understanding the disease mechanisms of COVID-19 and guide us in the prevention and treatment of COVID-19.

Core region typically consists of genes specific for the underlying disease. Based on the hypothesis that the similar molecular mechanism of diseases lies in their overlapped peripheries, we identify the molecular mechanism of disease causation, new comorbidity and aid rational drug target for COVID-19. In particular, we construct the flower model for COVID-19, SARS and H1N1 and show the details of their overlapped peripheral proteins. Enrichment analysis further proved that overlapped peripheries consistently enrich in Parkinson Disease, Huntington Disease, Alzheimer Disease and Non-Alcoholic Fatty Liver Disease pathways; provide 16 proteins targeted by 9 existing drugs currently undergoing clinical trials and drugs predicted by periphery have a certain effectiveness in the treatment of COVID-19. The periphery and core structure of COVID-19 provides new insights for the analysis of disease relationship and drug prediction.

Although Ratnakumar *et al.* [21] proposed instructive methods for identifying disease core genes, there are variety of definitions including highest differential expression level, strongest effect mutations or interpretable mechanistic links to disease. As another reference, Sharma *et al.* [16] identified the core genes of asthma and represented a consensus list of genes collected based on their known association with asthma-related phenotypes, asthma-related pathology, OMIM, Gene to MeSH relationship, GWAS data and their network neighborhood. The biggest hits from GWAS have helped pinpoint important core genes, but there still have lower frequency variants of larger effects. Quantification of disease-causing effects and identification of core remains open questions.

Generally, disease is driven by an accumulation of weak effects on the key genes and regulatory pathways that drive disease risk. Liu *et al.* [20] interpreted disease in a paradigm, in which the effects of weak trans-eQTL SNPs are accumulated and mediated through peripheral genes to impact the expression of core genes. The weak effects of variation in peripheral gene can be amplified by regulating core genes. Topological characteristics of VHN and core region have proved this from another perspective. VHN forms an inwardly compact module (high cohesiveness), indicating that the weak effects of peripheral variation gather in disease neighborhood. Instead, core region forms a stretched subnetwork (reduced cohesiveness) in VHN, indicating that core interacts with the wider peripheral region and receives more signals of weak effects. The flower model typically shows how effects of variation in common peripheral genes influence different diseases by mediating into different cores. Mining accumulation and mediation graph pattern of peripheral variation will be next key steps in deciphering disease.

Materials and methods

Materials for model building

Human interactome

Human interactome is from the underlying network using the experimentally documented molecular interactions in human cells from the interactome platform [15]. The data contains 16 461 proteins and 239 305 physical interactions (details in Table S1 available online at https://github.com/wangbingbo2019/ENCORE-of-COVID-19); several sources of protein interactions are combined: (i) Binary interactions from two available high-quality yeast-to-hybrid datasets; (ii) Literature curated interactions obtained by low throughput experiments; (iii) Kinase-substrate pairs and (iv) Signaling interactions.

SARS-CoV-2 host proteins

Gordon *et al.* [28] have produced the first systematic analysis of which human proteins SARS-CoV-2 may interact with during infection. Almost all SARS-CoV-2 viral genes are cloned and expressed in human HEK293T cells as 2xStrep-tag fusion proteins. The 29 tagged viral proteins are analyzed with affinity purification-mass spectrometry (AP-MS). They isolated them from lysates and systematically explored the host dependencies of the SARS-CoV-2 virus to identify host proteins already targeted with existing drugs. They analyzed a total of 2750 human proteins, and in these proteins, high confidence VHPs were identified using SAINTexpress [29] and the MIST algorithm [31, 32]. Finally, they discovered 332 high confidence proteins interacting with SARS-CoV-2 viral genes.

SAINT_BFDR and MIST score

Significance Analysis of INTeractome (SAINT) is a statistical method for probabilistically scoring protein–protein interaction data from AP-MS experiments. Teo *et al.* [29] presented a new implementation, SAINTexpress, an upgraded implementation of SAINT for filtering high confidence interaction data from AP-MS experiments. SAINTexpress reports the Bayesian False Discovery Rate (BFDR) estimates at all probability thresholds, which is computed directly from the posterior probabilities as

$$\mathrm{BFDR}\left(p^*\right) = \frac{\sum_{ij}\left(1-p_{ij}\right) I\left\{p_{ij} > p^*\right\}}{\sum_{ij} I\left\{p_{ij} > p^*\right\}}, \tag{1}$$

where $I\{A\}$ denotes the indicator function of event A. With this information, the user can determine the probability thresholds to control the BFDR at the target rate.

AP-MS experiments can identify a large number of protein interactions, but only a fraction of these interactions are biologically relevant. Verschueren *et al.* [32] described a comprehensive computational strategy to process raw AP-MS data, performed quality controls and prioritized biologically relevant bait-prey pairs in a set of replicated AP-MS experiments with Mass spectrometry Interaction STatistics (MIST). The MIST score is a linear combination of prey quantity (abundance), abundance invariability across repeated experiments (reproducibility) and prey uniqueness relative to other baits (specificity). The MIST pipeline is implemented in R. The most recent version of the MIST pipeline can be downloaded from GitHub (https://github.com/everschueren/MIST).

Peripheral and core regions detection process

The detection of periphery region and core region is based on the local maxima of connectivity significance between the VHPs above the threshold. We use the VHPs with Saint_BFDR ≤ 0.05 in detection, a total of 1160 VHPs. Then at increasing MIST score cutoffs, 0.1, 0.5, 1.0, as different thresholds, we select the corresponding subset of VHPs and identify the size of the induced LCC ($S_{LCC_{VHPs}}$). And we compute the size of the LCC of the same number of random proteins 1000 times in the human interactome; we get 1000 $S_{LCC_{ran}}$, then LCC's z-score is given by

$$z-\text{score} = \frac{S_{LCC_{VHPs}} - \mu\left(S_{LCC_{ran}}\right)}{\sigma\left(S_{LCC_{ran}}\right)}, \quad (2)$$

where $\mu(S_{LCC_{ran}})$ represents the expected value, and $\sigma(S_{LCC_{ran}})$ represents standard deviation.

Then, we get a curve by connecting LCC's z-scores of increasing MIST scores. We identify two peaks in the curve of the z-score values as two local maxima of connectivity significance between the VHPs. At the first peak, the LCC of the corresponding VHPs subset is called VHN. And at the second peak, the LCC of the corresponding VHPs subset is called core region. The core region is removed from the VHN and what is left is the periphery region.

Topological properties analysis

Internal and external connectivity

To investigate whether VHPs tend to form inwardly compact module, we tested internal connectivity and external connectivity [85] of two protein sets of 78 core proteins and 1012 VHN proteins in the human interactome. For a protein set S, m_S is the number of edges between proteins in S, n_S is the number of proteins in S and c_S is the total number of edges leaving S, which is the edge where one node is inside S and the other node outside of S, n is the number of proteins in the whole human interactome.

Internal connectivity: Internal density: $f(S) = m_S/(n_S\,(n_S - 1)/2)$ is the internal edge density of the core proteins set S. Edges inside: $f(S) = m_S$ is the number of edges between the members of S. Average internal degree: $f(S) = 2m_S/n_S$ is the average internal degree of the members of S.

External connectivity: Expansion measures the number of edges per protein that point outside the protein set S: $f(S) = c_S/n_S$. Cut Ratio is the fraction of existing edges (out of all possible edges) leaving the protein set S: $f(S) = c_S/(n_S(n - n_S))$.

Combine internal and external connectivity: Conductance: $f(S) = c_S/(2m_S + c_S)$ measures the fraction of total edge volume that points outside the protein set S. Cohesiveness: $f(S) = m_S/(m_S + c_S)$ measures the fraction of total internal edge volume of the protein set S.

The significance of connectivity was quantified based on z-score: z-score $= (f(S) - \mu)/\sigma$, where μ and σ are the mean and variance of 1000 randomly connected components, respectively.

Generation of random connected component

To evaluate the structure of VHN or core region, we generated random connected components using the following procedure: we start with a random candidate protein set and by adding in each step a small number of proteins from the periphery, and we extend this set until it induces a LCC that has a similar cardinality as the core or VHN. The percentage of each expansion of the candidate protein set is set to 0.01 in our analysis.

Materials for comorbidity and drug repurposing analysis

Disease-gene associations

The corpus of 70 diseases is manually chosen by Ghiassian *et al.* [12], with the additional criteria of at least 20 associated genes reported in the literature for every disease. The disease-gene associations are retrieved from OMIM (http://www.ncbi.nlm.nih.gov/omim) [25] and GWAS (Genome-Wide Association Studies). The OMIM associations they use also include associations from UniProtKB/Swiss-Prot and have been compiled [86]. The disease-gene associations from GWAS are obtained from the PheGenI database (PhenotypeGenotype Integrator; http://www.ncbi.nlm.nih.gov/gap/PheGenI) [5] that integrates various NCBI genomic databases. They use a genome-wide significance cutoff of P-value $= 5 \times 10^{-8}$. In addition, we collect expression data GSE5972 and GSE27131 directly related to SARS and H1N1 from the Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo), respectively, then analyse differential expression of gene with GEO2R tool (https://www.ncbi.nlm.nih.gov/geo/geo2r/). In the end, we obtain a total of 2913 associated genes of 72 disease modules, including SARS and H1N1 (details in Table S2 available online at https://github.com/wangbingbo2019/ENCORE-of-COVID-19).

C3 algorithm

The Connect separate Connected Components (C3) algorithm [13] is a disease module detection algorithm based on the connectivity significance of nodes and edges in a network. Firstly, the connected components set of disease proteins is determined, and the direct neighbors of disease proteins are taken as candidate proteins. Then, the P-value based on hypergeometric distribution is used to calculate the connection probability of candidate proteins and candidate edges, so as to characterize the ability of candidate proteins in connecting the connected components of disease proteins. Finally, by using a greedy process to detect the intermediate proteins for connecting the connected components, a succinct disease module dominated by disease proteins is presented.

Disease similarity

Given two disease modules A and B, we define the average shortest distance between disease modules A and B

$$\langle d_{AB}\rangle = \frac{1}{|A| + |B|}\left(\sum_{a\in A}\min_{b\in B} d(a,b) + \sum_{b\in B}\min_{a\in A} d(a,b)\right), \quad (3)$$

where $d(a, b)$ represents the shortest path length between node a and b in the network and $|A|$ and $|B|$ represent the size of disease modules A and B, respectively.

The disease similarity between A and B is given by the following equation:

$$\text{sim}_{AB} = 1 - \frac{\langle d_{AB}\rangle}{\langle d\rangle_{\max}}, \quad (4)$$

where $\langle d\rangle_{\max}$ represents the maximum average shortest distance between all disease pairs. The value range of sim_{AB} is 0–1, and the closer to 0, the lower the disease similarity.

Mediator proteins detection

In order to find the mediator proteins between the two diseases, Maiorino *et al.* defined a topological measure called Flow Centrality (FC) [14], identifying the proteins that are involved in most of the molecular interactions occurring between the two disorders. FC: Given two disease modules A and B, the FC of a node m is given by $FC_{A,B}(m)$

$$FC_{A,B}(m) = \frac{1}{|A||B|} \sum_{a \in A, b \in B} \frac{s_{ab}(m)}{s_{ab}}, \tag{5}$$

where $s_{ab}(m)$ is the number of shortest paths from a to b passing through node m, s_{ab} is the total number of shortest paths between a and b and $||$ is the size of the corresponding set.

The statistical significance of the obtained values is calculated by comparing them with the random 1000 times module pairs. For each random pair of two modules, we calculate the FC of each node in the network and measure the average $\mu(FC_{ran})$ and standard deviation $\sigma(FC_{ran})$ across all the samples. The FCS of a node m is then calculated as

$$FCS_{A,B}(m) = \frac{FC_{A,B}(m) - \mu(FC_{ran})}{\sigma(FC_{ran})}. \tag{6}$$

A large positive FCS indicates that the node is more likely to occur in the shortest path connecting the two modules, while a small or negative value suggests that the node is not relevant to the chosen pair of modules.

FC paths: All the shortest paths connecting the disease module A and B, whose intermediate proteins have a FCS of 2 or greater. In this work, we select all proteins in the FC paths between disease module A and B as the mediator proteins between the two diseases.

Drug target interactions

Feixiong Cheng *et al.* collected high-quality physical drug target interactions [64] on FDA-approved or clinically investigational drugs from six commonly used data sources: the DrugBank database (v4.3) [87], the Therapeutic Target Database (TTD, v4.3.02) [88], the PharmGKB database (30 December 2015) [89], ChEMBL (v20, accessed in December 2015) [90], BindingDB (downloaded in December 2015) [91] and IUPHAR/BPS Guide to PHARMACOLOGY (downloaded in December 2015) [92]. In total, 15 051 drug target interactions connecting 4428 drugs and 2256 unique human targets are built. In the human interactome, we just get 4380 drugs and 2161 unique human targets (details in Table S7 available online at https://github.com/wangbingbo2019/ENCORE-of-COVID-19).

Network proximity

Given V, the set of VHPs, T, the set of drug targets, and $d(v,t)$ the shortest path length between node $v \in V$ and $t \in T$ in the network, we define $\langle d_{vt} \rangle$ according to Eq. (3) to quantify the network-based distance between VHPs and drug targets [27, 61, 62]. We determine the expected distances between two randomly selected sets of proteins, matching the size and degrees of the original V and T sets. The mean $\mu(d_{ran})$ and standard deviation $\sigma(d_{ran})$ of the reference distribution allow us to get the *z-score* of the distance $\langle d_{vt} \rangle$, defined as

$$z - score = \frac{\langle d_{vt} \rangle - \mu(d_{ran})}{\sigma(d_{ran})}. \tag{7}$$

The smaller the z-score, the closer the distance between the VHPs and drug targets in the network, which implies that the drug is more likely to perturb the disease.

Expression perturbation profiles

We obtain drug perturbation profiles from the Connectivity Map (CMap) database [81, 82] by using the Python package CMapPy [93]. For each perturbation profile, we calculate the significance of size of overlap between the perturbed genes ($|Z\text{-Score}| > 2$) and SARS-CoV-2 targets derived from Gordon *et al.* [28], using Fisher's Exact Test. We also retrieve gene expression data of the cell line A549 after infection with SARS-CoV-2 [84]. The Spearman correlation coefficient is employed in estimating the correlation between the perturbation scores provided in CMap and the gene expression fold change caused by SARS-CoV-2 infection.

ROC curve and AUC score

We use drug rankings to plot ROC curves and calculate AUC scores for performance analysis. The AUC score measures the quality of differentiating between positive and negative situations. For the sorted table, we use different z-scores as thresholds to calculate the FPR and the TPR to draw the ROC curve. AUC scores range from 0 to 1, with 1 representing complete performance and 0.5 representing the performance of the random classifier. We use the Python package scikit-learn [94] to plot ROC curves and calculate AUC scores.

Key Points

- Here, we use network medicine framework to uncover the peripheral and core regions of SARS-CoV-2 perturbed neighborhood in human interactome, and construct an omnigenic virus-host network (VHN) to study COVID-19 systematically.
- We find that peripheral region can be used to improve the results for identifying comorbidities as well as detecting drug repurposing candidates for COVID-19 based on network proximity with modules of other 72 well curated diseases.
- Furthermore, by identifying the overlapped peripheral region of COVID-19, SARS and H1N1 as a flower model, we present some common molecular mechanisms and drug targets for these diseases.
- Our study illustrates the potential application of omnigenic VHN including peripheral and core regions as a powerful pattern in prevention and treatment of COVID-19.

Supplementary data

Supplementary data are available online at https://github.com/wangbingbo2019/ENCORE-of-COVID-19.

Data Availability

The dataset used in this study, as described in the Materials and Methods paragraph, is available as Supplementary Data.

Acknowledgments

We would like to thank the developers of all tools mentioned in this paper. Without the software they developed, the

presented work could not exist. We also thank Menche *et al*. for the original human interactome network data, thank Dr. Chenxing Zhang for reviewing the manuscript and thank all reviewers for their helpful suggestions.

Funding

National Natural Science Foundation of China (61772395, 61873198, 61702396, 61702397); China Postdoctoral Science Foundation (2015M582620); Fundamental Research Funds for the Central Universities (JB190306); Shanghai Municipal Science and Technology Major Project (2018SHZDZX01); LCNBI and ZJLab.

Conflict of Interest

We declare that there is no conflict of interest regarding the publication of this article.

References

1. Chen N, Zhou M, Dong X, *et al*. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet* 2020;**395**:507–13.
2. Li Q, Guan X, Wu P, *et al*. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *N Engl J Med* 2020;**382**:1199–207.
3. Venkatesan K, Rual J-F, Vazquez A, *et al*. An empirical framework for binary interactome mapping. *Nat Methods* 2009;**6**:83–90.
4. Barabasi A-L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. *Nat Rev Genet* 2011;**12**:56–68.
5. Ramos EM, Hoffman D, Junkins HA, *et al*. Phenotype-genotype integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. *Eur J Hum Genet* 2014;**22**:144–47.
6. Zanzoni A, Soler-López M, Aloy P. A network medicine approach to human disease. *FEBS Lett* 2009;**583**:1759–65.
7. Schadt E. Molecular networks as sensors and drivers of common human diseases. *Nature* 2009;**461**:218–23.
8. Pawson T, Linding R. Network medicine. *FEBS Lett* 2008; **582**:1266–70.
9. Califano A, Butte A, Friend S, *et al*. Leveraging models of cell regulation and GWAS data in integrative network-based association studies. *Nat Genet* 2012;**44**:841–7.
10. Feldman I, Rzhetsky A, Vitkup D. Network properties of genes harboring inherited disease mutations. *Proc Natl Acad Sci USA* 2008;**105**:4323–8.
11. Xu J, Li Y. Discovering disease-genes by topological features in human protein-protein interaction network. *Bioinformatics* 2006;**22**:2800–5.
12. Ghiassian S, Menche J, Barabasi A-L. A DIseAse MOdule detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Comput Biol* 2015;**11**:e1004120.
13. Wang B, Hu J, Zhang C, *et al*. C3: connect separate connected components to form a succinct disease module. *BMC Bioinformatics* 2020;**21**:433.
14. Maiorino E, Baek S, Guo F, *et al*. Discovering the genes mediating the interactions between chronic respiratory diseases in the human interactome. *Nat Commun* 2020;**11**:811.
15. Menche J, Sharma A, Kitsak M, *et al*. Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science* 2015;**347**:1257601.
16. Sharma A, Menche J, Huang C, *et al*. A disease module in the interactome explains disease heterogeneity, drug response and captures novel pathways and genes in asthma. *Hum Mol Genet* 2015;**24**:3005–20.
17. Vinayagam A, Gibson TE, Lee H-J, *et al*. Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets. *Proc Natl Acad Sci USA* 2016;**113**:4976–81.
18. Boyle E, Li Y, Pritchard J. An expanded view of complex traits: from polygenic to omnigenic. *Cell* 2017;**169**:1177–86.
19. Wray N, Wijmenga C, Sullivan P, *et al*. Common disease is more complex than implied by the core gene omnigenic model. *Cell* 2018;**173**:1573–80.
20. Liu X, Li Y, Pritchard J. Trans effects on gene expression can drive omnigenic inheritance. *Cell* 2019;**177**:1022–1034.e6.
21. Ratnakumar A, Weinhold N, Mar J, *et al*. Protein-protein interactions uncover candidate 'core genes' within omnigenic disease networks. *PLoS Genet* 2020;**16**:e1008903.
22. Sinnott-Armstrong N, Naqvi S, Rivas M, *et al*. GWAS of three molecular traits highlights core genes and pathways alongside a highly polygenic background. *bioRxiv* 2020. doi: 10.1101/2020.04.20.051631 preprint: not peer reviewed.
23. Sabik OL, Calabrese GM, Taleghani E, *et al*. Identification of a core module for bone mineral density through the integration of a co-expression network and GWAS data. *Cell Rep* 2020;**32**:108145.
24. Wang B, Glass K, Röhl A, *et al*. The periphery and the core properties explain the omnigenic model in the human interactome. *bioRxiv* 2019;749358 August 29, 2019, doi 10.1101/749358, preprint: not peer reviewed.
25. Hamosh A, Scott AF, Amberger JS, *et al*. Online Mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 2005;**33**: D514–7.
26. Zhou Y, Hou Y, Shen J, *et al*. Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. *Cell Discov* 2020;**6**:14.
27. Gysi D, do Valle Í, Zitnik M, *et al*. Network medicine framework for identifying drug repurposing opportunities for COVID-19. 2020.
28. Gordon D, Jang G, Bouhaddou M, *et al*. A SARS-CoV-2-human protein-protein interaction map reveals drug targets and potential drug-repurposing. *bioRxiv Prepr Serv Biol* 2020; 2020.03.22.002386.
29. Teo G, Liu G, Zhang J, *et al*. SAINTexpress: improvements and additional features in significance analysis of INTeractome software. *J Proteomics* 2014;**100**:37–43.
30. Fehr A, Perlman S. Coronaviruses: an overview of their replication and pathogenesis. *Methods Mol Biol* 2015;**1282**:1–23.
31. Jäger S, Cimermancic P, Gulbahce N, *et al*. Global landscape of HIV-human protein complexes. *Nature* 2011;**481**:365–70.
32. Verschueren E, Dollen J, Cimermancic P, *et al*. Scoring large scale affinity purification mass spectrometry datasets with MIST. *Curr Protoc Bioinformatics* 2015;**49**:8.19.1–16.
33. Shannon P, Markiel A, Ozier O, *et al*. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;**13**:2498–504.

34. Mick E, Kamm J, Pisco AO, *et al*. Upper airway gene expression differentiates COVID-19 from other acute respiratory illnesses and reveals suppression of innate immune responses by SARS-CoV-2. *MedRxiv Prepr Serv Heal Sci* 2020 May 19, 2020. doi: 10.1101/2020.05.18.20105171 preprint: not peer reviewed.
35. Addeo A, Obeid M, Friedlaender A. COVID-19 and lung cancer: risks, mechanisms and treatment interactions. *J Immunother Cancer* 2020;**8**.
36. Rogado J, Pangua C, Serrano-Montero G, *et al*. Covid-19 and lung cancer: a greater fatality rate? *Lung Cancer* 2020;**146**:19–22.
37. Bansal M. Cardiovascular disease and COVID-19. *Diabetes Metab Syndr* 2020;**14**:247–50.
38. Bonow RO, Fonarow GC, O'Gara PT, *et al*. Association of coronavirus disease 2019 (COVID-19) with myocardial injury and mortality. *JAMA Cardiol* 2020;**5**:751–3.
39. Guo T, Fan Y, Chen M, *et al*. Cardiovascular implications of fatal outcomes of patients with coronavirus disease 2019 (COVID-19). *JAMA Cardiol* 2020;**5**:811–8.
40. Cilia R, Bonvegna S, Straccia G, *et al*. Effects of COVID-19 on Parkinson's disease clinical features: a community-based case-control study. *Mov Disord* 2020;**35**:1287–92.
41. Sulzer D, Antonini A, Leta V, *et al*. COVID-19 and possible links with Parkinson's disease and parkinsonism: from bench to bedside. *NPJ Park Dis* 2020;**6**:18.
42. Chaudhry ZL, Klenja D, Janjua N, *et al*. COVID-19 and Parkinson's disease: shared inflammatory pathways under oxidative stress. *Brain Sci* 2020;**10**:807.
43. Freeman L. A set of measures of centrality based on betweenness. *Sociometry* 1977;**40**:35–41.
44. Freeman LC. Centrality in social networks conceptual clarification. *Soc Networks* 1978;**1**:215–39.
45. Saramäki J, Kivelä M, Onnela J-P, *et al*. Generalizations of the clustering coefficient to weighted complex networks. *Phys Rev E* 2007;**75**:27105.
46. Cameron M, Ran L, Xu L, *et al*. Interferon-mediated immunopathological events are associated with atypical innate and adaptive immune responses in patients with severe acute respiratory syndrome. *J Virol* 2007;**81**:8692–706.
47. Berdal J, Mollnes T, Wæhre T, *et al*. Excessive innate immune response and mutant D222G/N in severe a (H1N1) pandemic influenza. *J Infect* 2011;**63**:308–16.
48. Barrett T, Wilhite SE, Ledoux P, *et al*. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res* 2012;**41**:D991–5.
49. GeneCards—Human Genes|Gene Database|Gene Search. http://www.genecards.org/.
50. Huang C, Wang Y, Li X, *et al*. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* 2020;**395**:497–506.
51. Wang D, Hu B, Hu C, *et al*. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China. *JAMA* 2020;**323**:1061–9.
52. Lin L, Lu L, Cao W, *et al*. Hypothesis for potential pathogenesis of SARS-CoV-2 infection——a review of immune changes in patients with viral pneumonia. *Emerg Microbes Infect* 2020;**9**:1–14.
53. Mao L, Jin H, Wang M, *et al*. Neurologic manifestations of hospitalized patients with coronavirus disease 2019 in Wuhan, China. *JAMA Neurol* 2020;**77**:683–90.
54. Yang Y, Peng F, Wang R, *et al*. The deadly coronaviruses: the 2003 SARS pandemic and the 2020 novel coronavirus epidemic in China. *J Autoimmun* 2020;**109**:102434.
55. Potdar AA, Dube S, Naito T, *et al*. Reduced expression of COVID-19 host receptor, ACE2 is associated with small bowel inflammation, more severe disease, and response to anti-TNF therapy in Crohn's disease. *MedRxiv Prepr Serv Heal Sci* 2020 April 23, 2020. doi: 10.1101/2020.04.19.20070995 preprint: not peer reviewed.
56. Papadavid E, Scarisbrick J, Ortiz Romero P, *et al*. Management of primary cutaneous lymphoma patients during COVID-19 pandemic: EORTC CLTF guidelines. *J Eur Acad Dermatol Venereol* 2020;**34**:1633–6.
57. Lim S, Bae JH, Kwon H-S, *et al*. COVID-19 and diabetes mellitus: from pathophysiology to clinical management. *Nat Rev Endocrinol* 2021;**17**:11–30.
58. Spihlman AP, Gadi N, Wu SC, *et al*. COVID-19 and systemic lupus erythematosus: focus on immune response and therapeutics. *Front Immunol* 2020;**11**:589474.
59. Barabasi A-L, Oltvai Z. Network biology: understanding the cell's functional organization. *Nat Rev Genet* 2004;**5**:101–13.
60. Guney E, Menche J, Vidal M, *et al*. Network-based in silico drug efficacy screening. *Nat Commun* 2016;**7**:10331.
61. Cheng F, Desai R, Handy D, *et al*. Network-based approach to prediction and population-based validation of in silico drug repurposing. *Nat Commun* 2018;**9**:2691.
62. Cheng F, Lu W, Liu C, *et al*. A genome-wide positioning systems network algorithm for in silico drug repurposing. *Nat Commun* 2019;**10**:3476.
63. Yıldırım MA, Goh K-I, Cusick ME, *et al*. Drug—target network. *Nat Biotechnol* 2007;**25**:1119–26.
64. Cheng F, Kovacs I, Barabasi A-L. Network-based prediction of drug combinations. *Nat Commun* 2019;**10**:1197.
65. Fawcett T. Introduction to ROC analysis. *Pattern Recognit Lett* 2006;**27**:861–74.
66. Kamburov A, Stelzl U, Lehrach H, *et al*. The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res* 2012;**41**.
67. Clark L, Kodadek T. The immune system and neuroinflammation as potential sources of blood-based biomarkers for Alzheimer's disease, Parkinson's disease, and Huntington's disease. *ACS Chem Nerosci* 2016;**7**:520–7.
68. Fazzini E, Fleming J, Fahn S. Cerebrospinal fluid antibodies to coronavirus in patients with Parkinson's disease. *Mov Disord* 1992;**7**:153–8.
69. Butler M, Barrientos R. The impact of nutrition on COVID-19 susceptibility and long-term consequences. *Brain Behav Immun* 2020;**87**:53–4.
70. Sadasivan S, Sharp B, Schultz-Cherry S, *et al*. Synergistic effects of influenza and 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) can be eliminated by the use of influenza therapeutics: experimental evidence for the multi-hit hypothesis. *NPJ Park Dis* 2017;**3**:18.
71. Zarin D, Tse T, Williams R, *et al*. The ClinicalTrials.gov results database—update and key issues. *N Engl J Med* 2011;**364**:852–60.
72. Li J, Fan J-G. Characteristics and mechanism of liver injury in 2019 coronavirus disease. *J Clin Transl Hepatol* 2020;**8**:1–5.
73. Garrido I, Liberal R, Macedo G. Review article: COVID-19 and liver disease—what we know on 1st May 2020. *Aliment Pharmacol Ther* 2020;**52**.
74. Morshedi R, Bettis D, Majid M, *et al*. Bilateral acute iris transillumination following systemic moxifloxacin for respiratory illness: report of two cases and review of the literature. *Ocul Immunol Inflamm* 2012;**20**:266–72.
75. Winkelmann B, Leinberger H, Hertrich F, *et al*. Acute and chronic effects of low dose almitrine bismesylate in the

treatment of chronic bronchitis and emphysema. *Eur J Med* 1992;**1**:469–81.
76. Doshi P, Heneghan C, Jefferson T. Oseltamivir for influenza. *Lancet* 2016;**387**:124.
77. Beaucourt S, Vignuzzi M. Ribavirin: a drug active against many viruses with multiple effects on virus replication and propagation. Molecular basis of ribavirin resistance. *Curr Opin Virol* 2014;**8C**:10–5.
78. Kvasnicka H, Thiele J, Bueso-Ramos C, *et al*. Long-term effects of ruxolitinib versus best available therapy on bone marrow fibrosis in patients with myelofibrosis. *J Hematol Oncol* 2018;**11**:42.
79. Critchley I, Jones M, Heinze P, *et al*. In vitro activity of levofloxacin against contemporary clinical isolates of legionella pneumophila, mycoplasma pneumoniae and chlamydia pneumoniae from North America and Europe. *Clin Microbiol Infect* 2002;**8**:214–21.
80. DeVore R, Hainsworth J, Greco F, *et al*. Chronic oral etoposide in the treatment of lung cancer. *Semin Oncol* 1993; **19**:28–35.
81. Subramanian A, Narayan R, Corsello S, *et al*. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* 2017;**171**:1437–1452.e17.
82. Lamb J, Crawford E, Peck D, *et al*. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 2006;**313**:1929–35.
83. Buhaescu I, Izzedine H, Covic A. Sirolimus—challenging current perspectives. *Ther Drug Monit* 2006;**28**(5): 577–84.
84. Blanco-Melo D, Nilsson-Payant BE, Liu W-C, *et al*. SARS-CoV-2 launches a unique transcriptional signature from in vitro, ex vivo, and in vivo systems. *bioRxiv* 2020 March 24, 2020. doi: 10.1101/2020.03.24.004655 Arxiv biorxiv;2020.03.24.004655v1, preprint: not peer reviewed.
85. Yang J, Leskovec J. Defining and evaluating network communities based on ground-truth. *Knowl Inf Syst* 2015;**42**:181–213.
86. Mottaz A, Yip Y, Ruch P, *et al*. Mapping proteins to disease terminologies: from UniProt to MeSH. *BMC Bioinformatics* 2008;**9**(Suppl 5):S3.
87. Law V, Knox C, Djoumbou Y, *et al*. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res* 2014;**42**:D1091–7.
88. Zhu F, Shi Z, Qin C, *et al*. Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic Acids Res* 2011;**40**:D1128–36.
89. Hernandez-Boussard T, Whirl-Carrillo M, Hebert J, *et al*. The pharmacogenetics and pharmacogenomics knowledge base: accentuating the knowledge. *Nucleic Acids Res* 2008;**36**:D913–8.
90. Gaulton A, Bellis L, Bento A, *et al*. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* 2011;**40**:D1100–7.
91. Liu T, Lin Y, Wen X, *et al*. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res* 2007;**35**:D198–201.
92. Pawson AJ, Sharman JL, Benson HE, *et al*. The IUPHAR/BPS guide to PHARMACOLOGY: an expert-driven knowledgebase of drug targets and their ligands. *Nucleic Acids Res* 2014;**42**:D1098–106.
93. Enache OM, Lahr DL, Natoli TE, *et al*. The GCTx format and cmap{Py, R, M, J} packages: resources for optimized storage and integrated traversal of annotated dense matrices. *Bioinformatics* 2019;**35**:1427–429.
94. Pedregosa F, Varoquaux G, Gramfort A, *et al*. Scikit-learn: machine learning in python. *J Mach Learn Res* 2011;**12**: 2825–830.
95. Waskom M, the seaborn development team. mwaskom/seaborn. 2020.