Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Medicine in Drug Discovery 10 (2021) 100090

Contents lists available at ScienceDirect

Medicine in Drug Discovery

journal homepage: www.elsevier.com/locate/medid

Research Paper

Network bioinformatics analysis provides insight into drug repurposing for COVID-19

Xu Li [a,*], Jinchao Yu [b], Zhiming Zhang [a], Jing Ren [a], Alex E. Peluffo [b], Wen Zhang [a], Yujie Zhao [a], Jiawei Wu [a], Kaijing Yan [a], Daniel Cohen [b], Wenjia Wang [a]

[a] GeneNet Pharmaceuticals, Tianjin, China
[b] Pharnext, Paris, France

ARTICLE INFO

Keywords:
COVID-19
Network bioinformatics
Drug repurposing
SARS-CoV-2
Network pharmacology

ABSTRACT

The COVID-19 disease caused by the SARS-CoV-2 virus is a health crisis worldwide. While developing novel drugs and vaccines is long, repurposing existing drugs against COVID-19 can yield treatments with known preclinical, pharmacokinetic, pharmacodynamic, and toxicity profiles, which can rapidly enter clinical trials. In this study, we present a novel network-based drug repurposing platform to identify candidates for the treatment of COVID-19. At the time of the initial outbreak, knowledge about SARS-CoV-2 was lacking, but based on its similarity with other viruses, we sought to identify repurposing candidates to be tested rapidly at the clinical or preclinical levels. We first analyzed the genome sequence of SARS-CoV-2 and confirmed SARS as the closest virus by genome similarity, followed by MERS and other human coronaviruses. Using text mining and database searches, we obtained 34 COVID-19-related genes to seed the construction of a molecular network where our module detection and drug prioritization algorithms identified 24 disease-related human pathways, five modules, and 78 drugs to repurpose. Based on clinical knowledge, we re-prioritized 30 potentially repurposable drugs against COVID-19 (including pseudoephedrine, andrographolide, chloroquine, abacavir, and thalidomide). Our work shows how in silico repurposing analyses can yield testable candidates to accelerate the response to novel disease outbreaks.

1. Introduction

The COVID-19 disease outbreak caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), formerly named “2019 novel coronavirus” (2019-nCoV), has already infected more than 108 million people and caused 2 million deaths in the world, in February 2021 [1]. While several vaccines have become available, the combat against the COVID-19 pandemic is still highly challenging because of the virus's emerging mutant strains, the difficulties of manufacturing and distributing vaccines, and more [2]. Another approach is small molecule drug research, especially drug repurposing approach [3], remains an important solution to find rapid therapies. At the time of the outbreak, being the time of this study, drug repurposing approach was one of the best strategies to explore efficient therapies against COVID-19 rapidly.

Drug repurposing can yield new therapies at a faster rate than novel drug discovery when the safety profiles of the drugs being repurposed have been evaluated in the context of drug development for another disease, and at an even faster rate when the drugs have been approved for other diseases and postmarketing safety surveillance data are available [4,5]. By relying on already known preclinical, pharmacokinetic, pharmacodynamic, and toxicity profiles of the drugs being repurposed, one can dramatically increase the rapidity of the response against a disease with unmet clinical needs, especially for an epidemic disease, where drug proven safe can be immediately tested. At the begging of the pandemic, in February 2020, more than 10 repurposed drugs were under clinical trials evaluation for COVID-19. Among them, Remdesivir (Gilead Sciences, in Phase 3, clinical trial No. NCT04257656), originally developed to treat the Ebola which showed inhibition of replicases in a broad range of viruses including coronaviruses, and chloroquine (in Phase 4 ChiCTR2000029975), originally approved as an antimalarial and autoimmune disease drug, which, unlike Remdesivir, doesn’t target viral proteins but works as human endosomal acidification fusion inhibitor, which may help to stop the virus’ infection lifecycle [6].

In silico methods offer a way to methodically and rapidly yield additional repurposing candidates [7]. For instance, when drug targets

* Corresponding author.
E-mail address: tsl-lixu@tasly.com (X. Li).

https://doi.org/10.1016/j.medidd.2021.100090
Received 26 January 2021; Revised 3 March 2021; Accepted 16 March 2021
Available online 30 March 2021

associated with the disease of interest are known, and when their protein structures or that of close homologs are available, it is possible to use structural bioinformatics to virtually screen (e.g., using molecular docking) a library of existing drugs against these known targets [8]. A study published on February 27, 2020, relied on this approach, using the predicted structure of all SARS-CoV-2 proteins based on their homology with other known coronavirus protein structures, and identified several compounds with potential antiviral activity [9].

Another approach to repurposing is the construction of so-called "disease-related molecular networks," i.e., interactions between gene products (sometimes together with cellular metabolites) involved in the etiology and symptoms of that disease [10]. There exist several ways to identify disease-related genes, whether using genomic data (e.g., Genome-Wide Association Studies), gene expression data (e.g., RNAseq differential expression analysis) or data directly collected from the scientific literature (e.g., text mining or expert curation, either analyzed in-house or via recognized structured databases). Compared to virtual screening, where the candidate targets are known from the start, network biology methods can identify additional, unanticipated targets, which are part of the same molecular pathways than previously known targets for the disease of interest [7,11].

In this study, we performed network bioinformatics analyses to repurpose existing drugs, which are at the completed Phase 2 stage or later, against the now pandemic COVID-19. At the time of the outbreak, our goal was to yield a list of experimentally testable repurposing drug candidates, despite the fact that little was known about SARS-CoV-2, by supplementing that little knowledge with extensive data on closely related viruses and machine-learning analysis of those data. Therefore, because in late January 2020, limited knowledge about COVID-19 was available, we focused our work on similar pathogens as indirect cues to identify COVID-19 related genes and build a molecular network that could serve the identification of repurposable drug targets. We first relied on genome sequence alignment of SARS-CoV-2 to identify SARS-CoV (Severe Acute Respiratory Syndrome Coronavirus) as the most similar virus, followed by MERS-CoV (Middle East Respiratory Syndrome Coronavirus) and other related human coronaviruses. We then applied our AutoSeed program, which performed text mining against all NCBI PubMed abstracts (referenced before January 2020) and systematic database research, which led to 34 COVID-19-related genes, including ACE2.

To study these disease genes and their role at the systems level, we used an iterative network-building algorithm "AutoNet" that expands, prunes and merges subnetworks, leading to a human COVID-19 disease network composed of 1344 genes. In total, 24 enriched pathways were identified in five topological network modules (i.e., community structure, a region where nodes are more densely connected, more likely to be related to the same function or disease [12]). We scanned this network for known drug-target interactions and applied proximity-based topology analysis [13] to obtain a list of 78 drugs repurposable against COVID-19. Finally, we manually filtered this list based on the criteria of the drugs' mechanisms of action, their adverse effects, and clinical approvals to yield a total of 30 drugs. In this study, we also discuss the repurposing and mechanisms of thalidomide in particular, since, after sharing our findings with multiple institutions and hospitals in China, one care unit reported the remission of a patient treated with this drug together with low-dose glucocorticoids. In addition, two clinical trials of thalidomide were registered.

2. Results

2.1. Genome sequence analysis suggests SARS as the most similar disease

After performing a BLASTn search using the SARS-CoV-2 (a.k.a. 2019-nCoV at the time of the analysis) genome sequence against the NCBI GenBank database (see Methods), representative sequences from top results, all being coronaviruses either in humans or other animals,

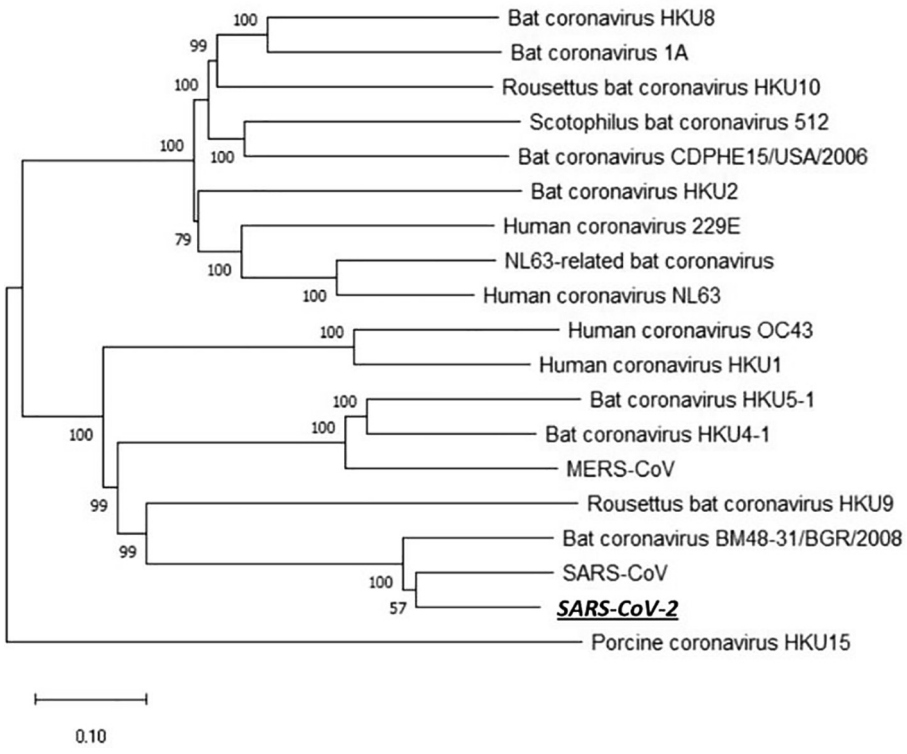

Fig. 1. Sequence analysis suggests SARS-CoV as the most similar virus to the SARS-CoV-2. Based on the results of BLASTn for SARS-Cov-2 against NCBI GenBank, nineteen genome sequences were selected as representative and were aligned using EMBI-EBI's MSA tool, and a neighbour-joining phylogenetic tree was built by the MEGA-X tool. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) is shown above the branches. The scale represents 0.10 residue substitutions per site.

were selected to build a phylogenetic tree using the neighbor-joining method (Fig. 1). We found SARS-CoV to be the evolutionarily closest sequence to SARS-CoV-2, with an 80% sequence identity. Among all other human coronaviruses, MERS-CoV is evolutionarily closest to SARS-CoV-2, with a 50% sequence identity. Importantly, we performed this analysis in January 2020, when the virus was less known and studied. Since then, multiple additional sequencing studies have been performed for SARS-CoV-2, including a landmark preprint, which suggested renaming 2019-nCoV to SARS-CoV-2 based on results similar to ours [14].

2.2. Text mining and database searches yield a list of 34 seed genes

In this step, we aimed to identify a list of human genes that are involved in the COVID-19 disease (Fig. 2A) and built a literature searching-engine-based web tool which is freely accessed in http://literature.tasly.com/covid19. Considering SARS-CoV as the closest virus to SARS-CoV-2, we used SARS as the first keyword for text mining against the database of NCBI PubMed. We searched for all human genes co-occurring with the keyword "SARS-COV-2" (abbreviations, full names, or synonyms) within any sentence (a.k.a "sentence co-occurrence" in NLP methods). We then ranked all genes based on their SARS co-occurrences count.

To enrich our text mining results, we added four other terms: "MERS", "coronavirus", "viral pneumonia", and "HIV" (Human Immunodeficiency Viruses). We chose MERS because of its close similarity to SARS-CoV-2 (Fig. 1) and the fact that it has been studied for long. "Coronavirus" and "viral pneumonia" were selected because they are highly related to the nature and symptoms of SARS and COVID-19, to the point that China and other regions of Asia, the synonyms of SARS and COVID-19 often contains the words "viral pneumonia". Although HIV does not belong to coronaviruses, "HIV" was used as a keyword because it was previously reported that HIV and SARS share similar viral protein structures [15] and that HIV drugs can be effective against SARS [16]. In addition, there exists extensive research and publication record on HIV, which can enrich our text mining analysis. For these four additional terms, the same co-occurrence analysis was performed, except that only the top 10% of each resulting list was retained. Therefore, the final text-mining-based list was made from the full SARS-related gene output list combined to these four top-10%-

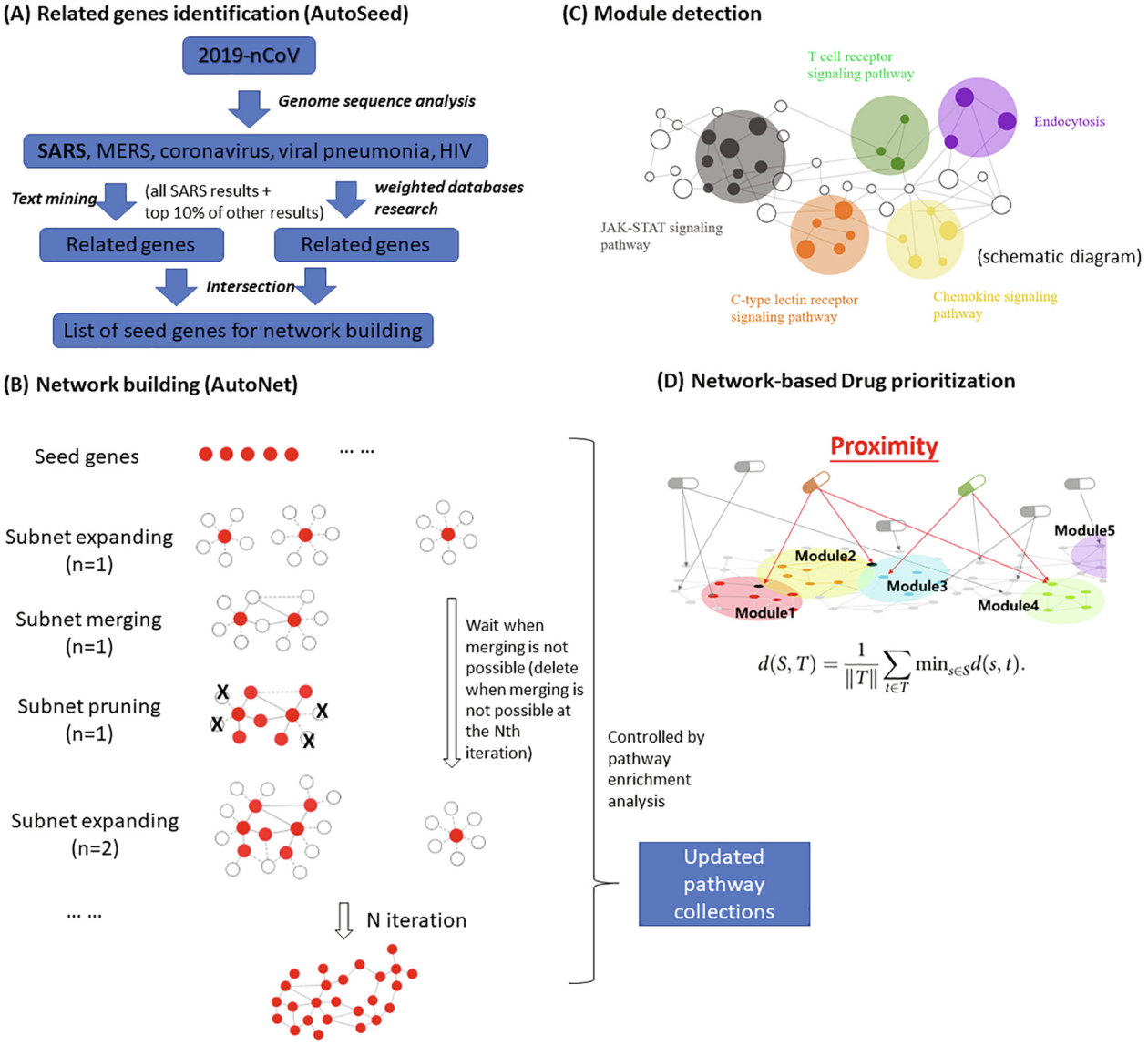

Fig. 2. The workflow of our network bioinformatics pipeline for SARS-CoV-2 drug repurposing. The equation shown in (D) represents how the distance was calculated to prioritize drugs based on proximity, see details in Methods.

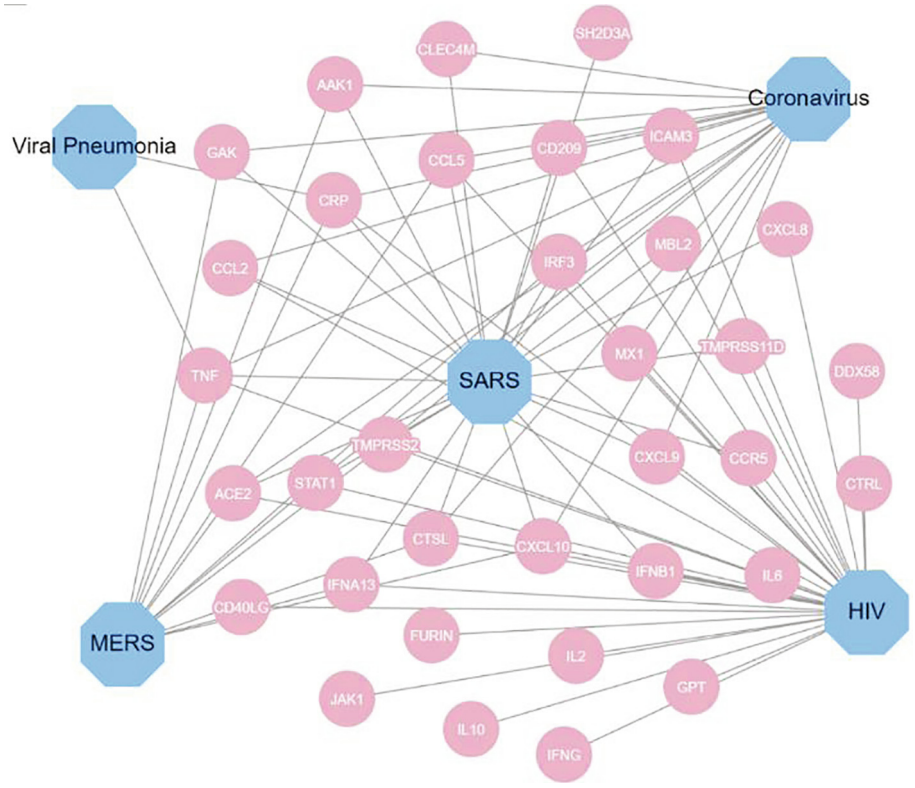

Fig. 3. Thirty-four genes related to SARS-CoV-2 identified by text mining and database searches. Each link represents at least one sentence co-occurrence in PubMed abstracts or at least one relationship recorded in one of our searched databases.

retained-gene lists (See Data availability for sources of extracted texts and papers).

In addition to our in-house text mining analysis, to enrich our search for SARS-related genes, we also searched for the five keywords aforementioned in reference databases including DisGeNET, DrugBank, KEGG, MalaCards, eDGAR, and GWAS-Catalog, because these databases integrate text-mining results with expert-curated information, from different aspects, including pathways, genetic factors, and animal models (see Methods).

A final list of seed genes was built by overlapping text-mining and database results (see Methods). This list contains 34 genes (shown as a network in Fig. 3, also see Data availability). Among them, 23 genes are directly linked to SARS. Two genes, CRP and TNF, connect to all keywords. Seven genes STAT1, CCL5, ACE2, IRF3, CXCL10, CTSL and TMPRSS2 are linked to four keywords (including SARS).

2.3. Network bioinformatics approach helps to predict 30 repurposable drugs

In order to contextualize and better understand, at a systems level, the molecular and physiological role of the COVID-19-related genes we found, we applied an in-house developed algorithm to build a molecular (i.e., protein) network taking these 34 genes as seeds. This algorithm repeats subnetwork expanding, merging, and pruning in an iterative manner, controlled by pathway enrichment analysis (see Fig. 2B and Methods). In this way, we obtained a final protein network of 1344 genes and 24 enriched pathways (see Data availability). The Newman greedy heuristic module detection algorithm was applied on the network, leading to five modules, representing the T cell receptor signaling pathway, JAK-STAT signaling pathway, C-type lectin receptor signaling pathway, Chemokine signaling pathway and Endocytosis (Fig. 2C). At last, DrugBank's drug-target interactions were added to the protein network, resulting a heterogeneous molecular network, over which proximity-based network analysis [13] identified a list of 78 repurposable drugs (see Data availability).

Having obtained these 78 drugs, we looked for more information, including clinical drug status, drug category, and adverse effect in the Yaozh (https://data.yaozh.com/) and DrugBank [17] databases. The former database provides increased China-related information,

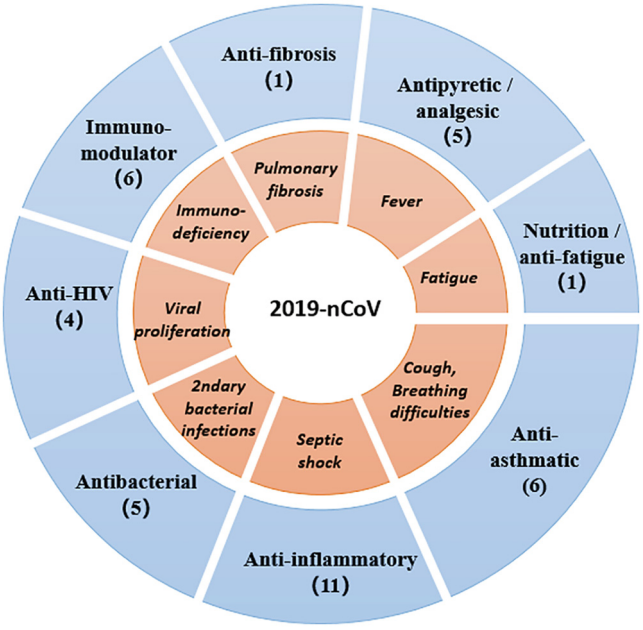

Fig. 4. Symptoms and mechanisms related to SARS-CoV-2 and the corresponding categories of our 30 suggested drugs.

including clinical trials in China, traditional Chinese medicine usage and theory, approvals by NMPA (National Medical Products Administration, formerly known as CFDA – China Food and Drug Administration) in China, and studies only published in Chinese, while the latter reports approval process by the U.S. FDA (Food and Drug Administration), known targets, therapeutic effects as well as basic chemical information [17].

Through a literature review, we identified a list of important symptoms and mechanisms linked to SARS-CoV-2, including fever, fatigue, cough [18], breathing difficulty, septic shock, viral proliferation, immunodeficiency and pulmonary fibrosis [19] (Fig. 4). We manually removed a drug from our list if it did not have any reported effect on any of these key symptoms and mechanisms. We also removed a drug from our list if it had strong reported side effects. We also filtered out drugs for which there is little scientific knowledge. After removing these drugs deemed unfit for rapid repurposing, we obtained a list of 30 drugs (Table 1).

2.4. Results sharing and case analysis

In order to help fight COVID-19 as fast as possible, we first publicly shared our list of 78 drugs (see Data availability) and our list of 24 enriched pathways (see Data availability) and we briefly explained our approach with healthcare professionals and hospitals, via GeneNet company's WeChat Chinese blog, on February 12, 2020. At the time, we put forward pseudoephedrine, andrographolide, chloroquine, abacavir, baricitinib, and quercetin as repurposing candidates from our list, because there were other researches also suggesting or pre-

Table 1
Thirty predicted drug candidates to repurpose against COVID-19. Type and Group were obtained by querying DrugBank. Initial ranks came from our proximity-based drug prioritization algorithm. Categories were obtained from Yaozh (a drug database in China), DrugBank and manual curation when the data was not available in neither of these databases.

DrugBank ID	Drug name	Type	Group	Initial rank	Category
DB00852	Pseudoephedrine	small molecule	approved	1	antipyretic or analgesic; antiasthmatic; anti-inflammatory
DB05767	Andrographolide	small molecule	investigational	2	antipyretic or analgesic; antiviral; anti-bacterial; anti-inflammatory
DB05513	Atiprimod	small molecule	investigational	3	immunomodulator
DB05017	YSIL6	small molecule	investigational	8	immunomodulator
DB06083	Tapinarof	small molecule	investigational	11	anti-inflammatory
DB00005	Etanercept	biotech drug	approved, investigational	12	antipyretic or analgesic; anti-inflammatory
DB00051	Adalimumab	biotech drug	approved	13	antipyretic or analgesic; anti-inflammatory
DB00065	Infliximab	biotech drug	approved	14	anti-inflammatory
DB00608	Chloroquine	small molecule	approved; investigational; vet_approved	15	anti-bacterial; anti-inflammatory
DB00668	Epinephrine	small molecule	approved; vet_approved	16	antiasthmatic
DB01041	Thalidomide	small molecule	approved; investigational; withdrawn	17	anti-fibrosis; immunomodulator
DB01407	Clenbuterol	small molecule	approved; investigational; vet_approved	18	antiasthmatic
DB01411	Pranlukast	small molecule	investigational	19	antiasthmatic
DB04956	Afelimomab	biotech drug	investigational	21	immunomodulator
DB06674	Golimumab	biotech drug	approved	32	antipyretic or analgesic; anti-inflammatory
DB09036	Siltuximab	biotech drug	approved, investigational	35	anti-viral
DB01250	Olsalazine	small molecule	approved	36	anti-inflammatory
DB12698	Ibalizumab	biotech drug	approved, investigational	39	anti-HIV
DB01327	Cefazolin	small molecule	approved	43	anti-bacterial
DB01048	Abacavir	small molecule	approved; investigational	50	anti-HIV
DB02375	Myricetin	small molecule	experimental	51	anti-inflammatory
DB04464	N-Formylmethionine	small molecule	experimental	52	immunomodulator
DB06475	Ruplizumab	biotech drug	Investigational	54	immunomodulator
DB00452	Framycetin	small molecule	approved	60	anti-bacterial
DB01009	Ketoprofen	small molecule	approved; vet_approved	65	anti-inflammatory
DB04835	Maraviroc	small molecule	approved; investigational	66	anti-HIV
DB06652	Vicriviroc	small molecule	investigational	69	anti-HIV
DB00172	Proline	small molecule	approved; nutraceutical	75	nutrition
DB04216	Quercetin	small molecule	experimental; investigational	76	antiasthmatic
DB11638	Artenimol	small molecule	experimental; investigational	78	anti-bacterial

dicting these drugs, mainly based on our Yaozh database search and literature review.

Chloroquine has been considered as one of the most promising repurposed drugs and is currently being tested against COVID-19 by more than ten clinical trials [6]. Abacavir was also predicted to treat COVID-19 by two separate studies [20]. Baricitinib was also suggested by the BenevolentAI company using their knowledge graph technology [21] and several clinical trials have been initiated (such as NCT04321993 in Phase 2). Finally, quercetin was predicted by a virtual screening studies of Chinese herbal medicines [22] and was later tested in clinical trial (NCT04377789) against COVID-19. The ongoing clinical trials and experimental validation for our 6 highlighted drugs, most of them beginning after our 1st result exchange, suggested that our approach succeeded in producing repurposing candidates that are worthy of further evaluation.

In a second exchange with partner experts from Chinese institutions and care units, via a webinar organized on February 22, we also put forward thalidomide as an interesting repurposing candidate as it was well ranked by our algorithm, the sole drug with anti-fibrosis effect in our list, while it was neither predicted nor tested by another research group. Later, successful use of thalidomide combined with low-dose glucocorticoid (methylprednisolone) was reported by a preprint for a 45-year old Chinese woman who had unsuccessfully been treated with ofloxacin (a fluoroquinolone antibiotic known to inhibit the DNA topoisomerase 4 subunit A and DNA gyrase subunit A of *Haemophilus influenzae*), oseltamivir (a.k.a Tamiflu, known to inhibit Neuraminidase of Influenza A virus) and lopinavir + ritonavir (a combination of antiviral drugs used to treat HIV known to target the HIV protein encoded by pol) (drug target information above from DrugBank [17]). We remind that these drugs do not belong to our proposed drugs as our method was repurposing drugs with human targets. Before being treated with thalidomide + methylprednisolone, the patient showed an increase in C-reactive protein (CRP) and cytokine levels, including interleukin 6 (IL-6), interleukin 10 (IL-10) and interferon-gamma (IFN-gamma) together with reduced CD4+ and CD8+ T cells counts. The authors reported that these abnormally high interleukin levels and abnormally low T cell levels returned to normal after three days of their combinatorial treatment.

It was previously shown that thalidomide enhances TCR (T cell receptor)-mediated T cells activation by by-passing T cell need for co-stimulation by accessory molecules, such as the B7 protein together with the CD28 protein, and therefore can overcome T cell deficiency [23]. In addition, previous work suggests that lenalidomide, a derivative of thalidomide, can restore T cells motility leading to their activation [24]. Finally, it was also reported that thalidomide prevents NF-kB from binding to the promoters of its target genes, including TNF-alpha and IFN-Gamma thereby reducing excessive inflammatory response [25,26]. Altogether, based on these previous studies, the reported successful use of thalidomide by Chen et al. [27], and our analysis, we hypothesize that thalidomide may be effective against COVID-19 by favorably modifying the immune response of the infected patients against the virus (Fig. 5). At the time of the preprint sharing (March 2020), thalidomide had been registered in two Phase 2 clinical trials: NCT04273529 and NCT04273581.

3. Discussion

We applied a network bioinformatics approach to prioritize potential drugs and their targets at the systems level based on pre-COVID-19 knowledge of related viruses. To our knowledge, until now, two other studies have investigated the COVID-19 disease using network-based repurposing. The first one took advantage of a knowledge graph (another type of network comprising different entity types, such as gene, protein, organism and disease, and relationship types, such as interacting with, phosphating, belonging to, etc.) technology to suggest baricitinib as potential treatment [21]. A second study used, in part, similar network techniques than reported in this study, although the main difference is that we relied on text-mining and database search for seed genes identification while they essentially relied on the use of transcriptomic data for enrichment analysis [28].

We would like to highlight that our network bioinformatics analysis relied not directly on the keyword COVID-19, but indirectly via its similar terms like SARS based on genome analysis and limited existing knowledge about the disease. This is because our study was conducted mainly in January and February 2020 when scientific knowledge of COVID-19 was seriously lacking. Now, almost one year after we first shared the preprint of this study, 30 out of the 34 genes we identified can now be found using the method of sentence cooccurrence with

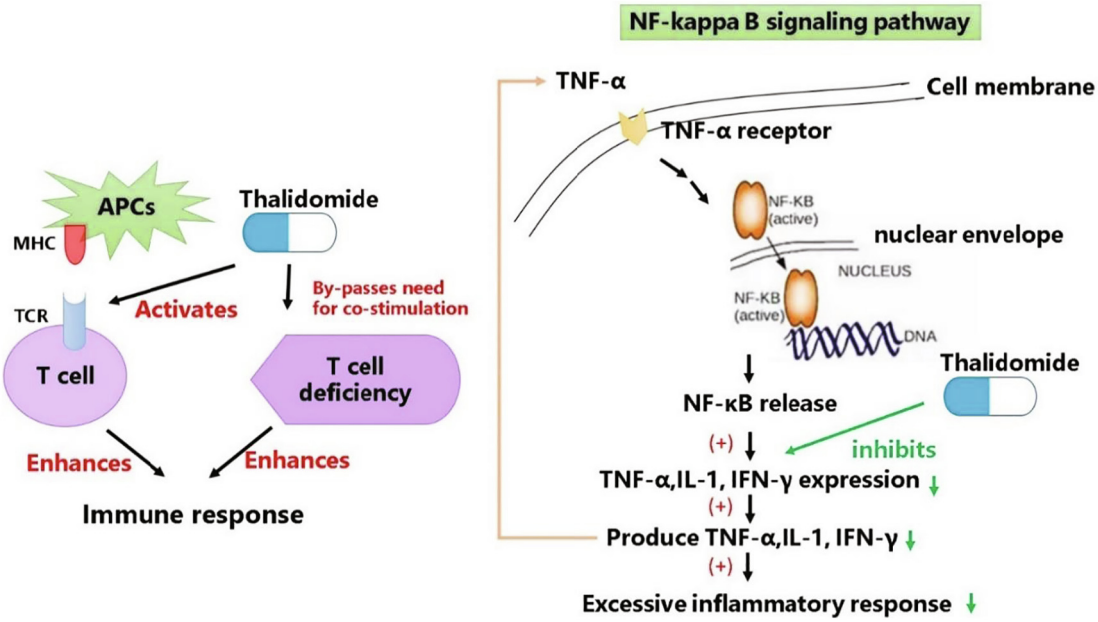

Fig. 5. Thalidomide's potential Mechanism-of-Action on COVID-19. APC: antigen-presenting cell; MHC: major histocompatibility complex; TCR: T cell receptor.

"Covid-19" based on the CORD-19 text dataset (version November 2020) [29] (see more proofs in Data Availability).

The purpose of this *in silico* work is not to yield repurposing candidates which should be immediately given to patients, rather, it is to shorten the immense list of candidates as to focus rigorous clinical (sometimes and experimental) evaluation on a smaller number of candidates. In the context of an outbreak, the response must be swift but must also satisfy the usual safety and quality standards of the medical community. Deriving a list of candidates based on the available robust in silico data and analysis allows the expert to concentrate scarce resources on evaluating a smaller number of options but with the same level of standards. We therefore designed our analyses hoping that our results could be helpful in rapidly designing and implementing clinical trials or preclinical experiments to treat COVID-19 considering the available preclinical, pharmacokinetic, pharmacodynamic, toxicity, and clinical knowledge. Extreme caution is needed for drugs with important side effects, even when they are already approved drugs, because the interactions of the side effects with the new disease are unknown. In such a situation, combinations are an interesting path, as they can be more efficient at lower synergistic doses when used synergistically while suppressing their side effects [7]. As of now, 7 of the 30 proposed drugs have been tested in clinical trials according to clinicaltrials.gov database (search in February 2021, see more details in Data Availability). To our knowledge, chloroquine (or hydroxychloroquine) is the only drug with published results on clinical trials, which mostly suggest a lack of efficacy in COVID-19 patients and safety concerns at high doses [30]. We believe that drug repurposing, preferably coupled with synergetic combinations at low doses, could help find other therapies in addition to recent successful vaccine development against COVID-19.

4. Materials and methods

4.1. Genome sequence analysis

From NCBI GenBank, the complete genome of Wuhan-Hu-1 (NC_045512.2) was downloaded as the 2019-nCoV sequence. This genome sequence was used to search for closely related viruses, against the whole database using BLASTn (default parameters except that we obtained more results than 100 by default). Among the BLASTn results, we extracted the following complete genome sequences as representative to build a phylogenetic tree: SARS coronavirus (SARS-CoV), MERS coronavirus (MERS-CoV), Human coronaviruses OC43, NL63, HKU1 and 229E; Bat coronaviruses BM48-31/BGR/2008, CDPHE15/USA/2006, HKU8, HKU5-1, 1A, HKU4-1 and HKU2; Rousettus bat coronaviruses HKU9 and HKU10; NL63-related bat coronavirus strain BtKYNL63-9a; Scotophilus bat coronavirus 512; Porcine coronavirus HKU15 (see the full table with their genome identifiers in Data Availability). Multiple sequence alignment was calculated by EMBL-EBI's MSA (multiple sequence alignment) tool (https://www.ebi.ac.uk/Tools/msa/) using default parameters. A tree was built using the neighbor-joining method with the MEGA-X software [31], using the maximum composite likelihood model and 1000 bootstraps. The resulting tree was represented using the phylogram format (i.e., a tree branch lengths are proportional to the amount of inferred evolutionary change) [32].

4.2. Related genes identification

PubMed (version 2019-12) was downloaded from its FTP site. Note that no article mentioning SARS-CoV-2 (or its previous name 2019-nCoV) had been published before that date, meaning that our text mining analysis did not directly consider the COVID-19 disease. Instead, it aims at predicting the network base on closely related viruses and their physiology. More than 29 million abstracts were processed for sentence and word tokenization by the natural language processing tool Spacy (v2). Inputted keywords of interest (SARS, MERS, coronavirus, viral pneumonia, and HIV) were extracted by exact matches to detect abbreviations or regression expressions to detect full names or synonyms. Entity recognition for genes was proceeded by mapping gene names and unambiguous synonyms from the HGNC database. Co-occurrence numbers were counted by the number of papers where a pair of gene and an input entity was in one sentence. A list of related genes ranked by sentence co-occurrence numbers was obtained for each of the five input entities. The final text-mining resulting list (the network shown in Fig. 2) was built from the whole list for SARS and the top 10% of each of the other four lists.

Database search for related genes was performed by a program developed in-house, AutoSeed, which can search for disease-related genes in the following databases: DisGeNET [33], DrugBank [17], KEGG [34], Malacards [35], eDGAR [36], NHGRI-EBI GWAS-Catalog [37]. Note that this program was developed for all types of diseases, and not specifically for viral diseases. Its function is to interrogate all of these databases automatically and to return a list of related genes sorted by the number of times they occur in those databases. Although the GWAS-Catalog is one of the resources of AutoSeed, for SARS and MERS, because there are no published GWAS, the findings in that category are, as expected, null. The final database-based list was composed of the whole list for SARS and the top 10% of each of the other four lists.

4.3. Network building

Network building was performed automatically by another of our in-house program "AutoNet", implemented on our drug discovery cloud platform (CloudPhar: http://cloud.tasly.com/#/portalHome). This algorithm is explained by a schematic diagram in Fig. 2B.

Data for this step includes a local meta-pathway database for pathway enrichment analysis and a meta-PPI (protein–protein interaction) database to grow the network. The meta-pathway database is made of human pathways in KEGG [34] and Reactome (v70) [38] databases, after removing small pathways (less than five genes) and pathways which enrich too easily, such as hsa05200: Pathways in cancer. The meta-PPI database is composed of protein–protein interaction databases HPRD [39], BioGrid [40] (excluding genetic interactions), and STRING [41] (excluding PPIs with confidence score <0.7).

The building process repeats network expanding, merging, and pruning in an iterative manner. At the initial state, all seed genes are considered as positive nodes where each seed gene is a subnetwork (i.e., connected component) composed of one node. A dynamic pathway collection for network building is initiated by a pathway enrichment analysis (hypergeometric test, False Discovery Rate correction, threshold: adjusted p-value <0.001) for all positive nodes against our meta-pathway database. Here, a subnetwork is used to denote any growing network during the network building process and to be distinguished from our final network; a pathway means any pathway from meta-pathway databases (KEGG and Reactome). In each expanding step, protein interactors of any positive nodes are added as temporary nodes according to our meta-PPI database. In each merging step, only the pair of subnetworks that share the most positive nodes and temporary nodes are merged, while the other subnetworks wait to be merged in the next iterations. In the subnetwork pruning step, those temporary nodes which are not in any of the pathway collection in the current state are removed. Remaining nodes become positive nodes, and the dynamic pathway collection is updated by using pathway enrichment analysis for all positive nodes.

Sub-networks are grown until they cannot be further merged. At last, if more than one subnetwork remains, only the largest connected

component and any other subnetwork whose size is greater than 5% of largest connected component's size, are kept. In this study, only the largest component was kept because the others were too small.

4.4. Network-based drug repurposing

After the network was built, core modules were detected (Fig. 2C), using the Newman greedy heuristic algorithm [42], implemented in igraph package (v1.2.4.2) in the R language (version 3.5.3). Potential drugs were then mapped to the COVID-19 network through drug-target interactions (source from DrugBank). As shown in Fig. 2D, different drugs can be linked to one or more different modules (shown as colored areas) in the network. In order to find the maximum effective coverage of the core functional modules for each drug, we used a proximity method with each drug proximity distance calculated as the mean value of the shortest distances between any drug and each of the core modules in the space (equation shown in Fig. 2D) [13].

5. Conclusion

In this study, we applied a network bioinformatics approach to repurpose drugs for COVID-19. Our seed genes (i.e., disease-related genes) resulted from our AutoSeed program -- a systematic text mining and database search, while our protein network was built by AutoNet, mainly based on knowledge of pathways, protein–protein interaction and graph theory. Combining these results with module detection and proximity analysis algorithms allowed us to identify 78 old drugs repurposable for COVID-19 disease. Finally, drug database search and manual curation helped shorten our first list to a final list of 30 rapidly repurposable drugs to be tested clinically and experimentally, possibly as combination therapies to treat COVID-19 patients.

Conflict of interest

X.L., Z.Z., J.R., W.Z., Y.Z., J.W., K.Y. and W.W. are employees of GeneNet Pharmaceuticals. J.Y., A.E.P. and D.C. are employees of Pharnext.

Funding

This research received no external funding.

CRediT authorship contribution statement

Xu Li: Conceptualization, Formal analysis, Investigation, Methodology, Validation, Writing - review & editing. **Jinchao Yu:** Formal analysis, Investigation, Writing - original draft, Writing - review & editing. **Zhiming Zhang:** Formal analysis, Investigation, Writing - review & editing. **Jing Ren:** Formal analysis, Investigation, Writing - review & editing. **Alex E. Peluffo:** Writing - original draft, Writing - review & editing. **Wen Zhang:** Investigation, Writing - review & editing. **Yujie Zhao:** Investigation, Writing - review & editing. **Jiawei Wu:** Investigation, Writing - review & editing. **Kaijing Yan:** Writing - review & editing. **Daniel Cohen:** Writing - review & editing. **Wenjia Wang:** Supervision, Writing - review & editing.

Acknowledgments

We would like to thank Zu Liu, and Yunhui Hu, Jia Sun, Hairong Wang from GeneNet, Serguei Nabirotchkin from Pharnext for fruitful discussions, and thank Zu Liu for his help in genome analysis.

Data Availability Statement

The datasets, including our codes and generated data, for this study can be found in the GitHub COVID-19-AIDrug repository (https://github.com/TaslyGeneNet/COVID19-AIDrug). Supplementary tables are in "article_supplementary_info" folder, containing (1) seed genes used in our algorithm, (2) enriched pathways, (3) the 78 resulting drugs to be repurposed, (4) information of genome sequences used in our sequence analysis, (5) SARS-CoV-2 seeds follow-up analysis, (6) follow-up analysis on recent clinical trials about the 30 repurposed drugs.

References

[1] WHO, COVID-19 Weekly Epidemiological Update. https://www.who.int/publications/m/item/weekly-epidemiological-update—16-february-2021. 2021.
[2] Kim JH, Marks F, Clemens JD. Looking beyond COVID-19 vaccine phase 3 trials. Nat Med 2021;27:205–11. https://doi.org/10.1038/s41591-021-01230-y.
[3] Wang X, Guan Y. COVID-19 drug repurposing: A review of computational screening methods, clinical trials, and protein interaction assays. Med Res Rev 2021;41:5–28. https://doi.org/10.1002/med.v41.110.1002/med.21728.
[4] Ashburn TT, Thor KB. Drug repositioning: Identifying and developing new uses for existing drugs. Nat Rev Drug Discov 2004;3:673–83. https://doi.org/10.1038/nrd1468.
[5] Pushpakom S, Iorio F, Eyers PA, Escott KJ, Hopper S, Wells A, et al. Drug repurposing: progress, challenges and recommendations. Nat Rev Drug Discov 2019;18:41–58. https://doi.org/10.1038/nrd.2018.168.
[6] Harrison C. Coronavirus puts drug repurposing on the fast track. Nat Biotechnol 2020;38:379–81. https://doi.org/10.1038/d41587-020-00003-1.
[7] Nabirotchkin S, Peluffo AE, Rinaudo P, Yu J, Hajj R, Cohen D. Next-generation drug repurposing using human genetics and network biology. Curr Opin Pharmacol 2020;51:78–92. https://doi.org/10.1016/j.coph.2019.12.004.
[8] Glaab E. Building a virtual ligand screening pipeline using free software: A survey. Brief Bioinform 2016;17:352–66. https://doi.org/10.1093/bib/bbv037.
[9] Wu C, Liu Y, Yang Y, Zhang P, Zhong W, Wang Y, et al. Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. Acta Pharm Sin B 2020;10:766–88. https://doi.org/10.1016/j.apsb.2020.02.008.
[10] Barabási AL, Gulbahce N, Loscalzo J. Network medicine: A network-based approach to human disease. Nat Rev Genet 2011;12:56–68. https://doi.org/10.1038/nrg2918.
[11] Hopkins AL. Network pharmacology. Nat. Biotechnol 2007;25:1110–1. https://doi.org/10.1038/nbt1007-1110.
[12] Newman MEJ. Modularity and community structure in networks. Proc Natl Acad Sci U S A 2006;103:8577–82. https://doi.org/10.1073/pnas.0601602103.
[13] Guney E, Menche J, Vidal M, Barábasi AL. Network-based in silico drug efficacy screening. Nat Commun 2016;7:1–13. https://doi.org/10.1038/ncomms10331.
[14] Gorbalenya AE, Baker SC, Baric RS, De Groot RJ, Gulyaeva AA, Haagmans BL, Lauber C, Leontovich AM. Severe acute respiratory syndrome-related coronavirus: The species and its viruses – a statement of the Coronavirus Study Group. BioRxiv 2020:1–15. https://doi.org/10.1101/2020.02.07.937862.
[15] Kliger Y, Levanon EY. Cloaked similarity between HIV-1 and SARS-CoV suggests an anti-SARS strategy. BMC Microbiol 2003;3:1–7. https://doi.org/10.1186/1471-2180-3-20.
[16] Yamamoto N, Yang R, Yoshinaka Y, Amari S, Nakano T, Cinatl J, et al. HIV protease inhibitor nelfinavir inhibits replication of SARS-associated coronavirus. Biochem Biophys Res Commun 2004;318:719–25. https://doi.org/10.1016/j.bbrc.2004.04.083.
[17] Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, MacIejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, Wilson M. DrugBank 5.0: A major update to the DrugBank database for 2018. Nucleic Acids Res 2018;46:D1074–82. https://doi.org/10.1093/nar/gkx1037.
[18] Xu Y-H, Dong J-H, An W, Lv X-Y, Yin X-P, Zhang J-Z, Dong L, Ma X, Zhang H-J, Gao B-L. Clinical and computed tomographic imaging features of Novel Coronavirus Pneumonia caused by SARS-CoV-2. J Infect 2020. https://doi.org/10.1016/j.jinf.2020.02.017.
[19] Shi H, Han X, Jiang N, Cao Y, Alwalid O, Gu J, et al. Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study. Lancet Infect Dis 2020;20:425–34. https://doi.org/10.1016/S1473-3099(20)30086-4.
[20] Beck BR, Shin B, Choi Y, Park S, Kang K. Predicting commercially available antiviral drugs that may act on the novel coronavirus (2019-nCoV), Wuhan, China through a drug-target interaction deep learning model. BioRxiv 2020. https://doi.org/10.1101/2020.01.31.929547.
[21] Richardson P, Griffin I, Tucker C, Smith D, Oechsle O, Phelan A, et al. Baricitinib as potential treatment for 2019-nCoV acute respiratory disease. Lancet (London, England) 2020;395:e30–1. https://doi.org/10.1016/S0140-6736(20)30304-4.
[22] Zhang D, Wu K, Zhang X, Deng S, Peng B. In silico screening of Chinese herbal medicines with the potential to directly inhibit 2019 novel coronavirus. J Integr Med 2020. https://doi.org/10.1016/j.joim.2020.02.005.

[23] Bartlett JB, Dredge K, Dalgleish AG. The evolution of thalidomide and its IMiD derivatives as anticancer agents. Nat Rev Cancer 2004;4:314–22. https://doi.org/10.1038/nrc1323.
[24] Riches JC, Gribben JG. Immunomodulation and immune reconstitution in chronic lymphocytic leukemia. Semin Hematol 2014;51:228–34. https://doi.org/10.1053/j.seminhematol.2014.05.006.
[25] Keddie S, Bharambe V, Jayakumar A, Shah A, Sanchez V, Adams A, et al. Clinical perspectives into the use of thalidomide for central nervous system tuberculosis. Eur J Neurol 2018;25:1345–51. https://doi.org/10.1111/ene.13732.
[26] Wen H, Ma H, Cai Q, Lin S, Lei X, He B, et al. Recurrent ECSIT mutation encoding V140A triggers hyperinflammation and promotes hemophagocytic syndrome in extranodal NK/T cell lymphoma. Nat Med 2018;24:154–64. https://doi.org/10.1038/nm.4456.
[27] Chen C, Qi F, Shi K, Yi L, Li J, Chen Y, et al., Thalidomide combined with low-dose glucocorticoid in the treatment of COVID-19 pneumonia, (2020) 1–6. https://www.preprints.org/manuscript/202002.0395/v1.
[28] Zhou Y, Hou Y, Shen J, Huang Y, Martin W, Cheng F. Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. Cell Discov 2020;6. https://doi.org/10.1038/s41421-020-0153-3.
[29] Wang LL, Lo K, Chandrasekhar Y, Reas R, Yang J, Eide D, et al. The COVID-19 open research dataset. ArXiv 2020.
[30] Kashour Z, Riaz M, Garbati MA, AlDosary O, Tlayjeh H, Gerberi D, et al. Efficacy of chloroquine or hydroxychloroquine in COVID-19 patients: a systematic review and meta-analysis. J Antimicrob Chemother 2021;76:30–42. https://doi.org/10.1093/jac/dkaa403.
[31] Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol 2018;35:1547–9. https://doi.org/10.1093/molbev/msy096.
[32] Pavlopoulos GA, Soldatos TG, Barbosa-Silva A, Schneider R. A reference guide for tree analysis and visualization. BioData Min 2010;3:1. https://doi.org/10.1186/1756-0381-3-1.
[33] Piñero J, Bravo Á, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, et al. DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res 2017;45:D833–9. https://doi.org/10.1093/nar/gkw943.
[34] Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M. New approach for understanding genome variations in KEGG. Nucleic Acids Res 2019;47:D590–5. https://doi.org/10.1093/nar/gky962.
[35] Rappaport N, Twik M, Plaschkes I, Nudel R, Stein TI, Levitt J, et al. MalaCards: An amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. Nucleic Acids Res 2017;45:D877–87. https://doi.org/10.1093/nar/gkw1012.
[36] Babbi G, Martelli PL, Profiti G, Bovo S, Savojardo C, Casadio R. eDGAR: A database of disease-gene associations with annotated relationships among genes. BMC Genomics 2017;18(S5). https://doi.org/10.1186/s12864-017-3911-3.
[37] Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res 2019;47:D1005–12. https://doi.org/10.1093/nar/gky1120.
[38] Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The Reactome pathway knowledgebase. Nucleic Acids Res 2018;46:D649–55. https://doi.org/10.1093/nar/gkx1132.
[39] Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human protein reference database - update. Nucleic Acids Res 2009;37(2009):767–72. https://doi.org/10.1093/nar/gkn892.
[40] Oughtred R, Stark C, Breitkreutz BJ, Rust J, Boucher L, Chang C, et al. The BioGRID interaction database: 2019 update, Nucleic Acids Res. 47 (2019) D529–D541. https://doi.org/10.1093/nar/gky1079.
[41] Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, et al. The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res 2017;45:D362–8. https://doi.org/10.1093/nar/gkw937.
[42] Newman ME. Fast algorithm for detecting community structure in networks. Phys Rev E 2004;69:066133.