COVID-19 Virus-Host Interactome of Lung Tissue PPI Analysis to Understand Biological Relevance and Therapeutic Modalities

BE 700 A1: AI in Systems Biology
Final Report
12/14/2020

Team Members:
Divya Venkat
Aishwarya Deengar
Nicole Carr
Tian Wang

INTRODUCTION

It is a well-established fact that viruses interact with host cellular pathways in order to replicate and evade immune responses. COVID-19 is caused by SARS-CoV-2 virus presented many initial challenges to determining these pathways due to its rapid transmission and evolution [1]. Transmission occurs primarily through respiratory droplets, and infection can manifest in asymptomatic carriers as well as patients with organ sepsis and acute respiratory failure. Some successful therapies at the time of this report include supportive management for acute hypoxic respiratory failure, dexamethasone to reduce 28-day mortality, and remdesivir to improve recovery time [1]. Research into possible treatment methods for the more severe symptoms is vital until the recently discovered vaccinations have been distributed to the general public.

Recent researches have identified the human proteins that interact with the SARS-CoV-2 referred to as Krogan proteins in this report [2]. Combined with protein-protein interaction (PPI) networks in human tissue and drug-target dataset, researchers have implemented different network propagation techniques to discover additional human proteins that interact with the SARS-CoV-2 and explore potential drug target sites. Efforts have also been made to identify repurposable drugs and potential drug combinations to combat this infection [3, 4].

Many research papers and post-mortem reports have confirmed damages ranging from moderate to severe in lung and brain tissues of COVID patients. This project aims to identify enriched protein complexes in lung tissue using the tissue PPI network and Krogan proteins that have a close interaction with SARS-CoV-2 so as to identify potential drug targets.

METHODS

The major workflow is described in figure.1. The first step was to filter protein complexes that involve at least one Krogan protein (human proteins directly interacting with SARS-CoV-2). These protein complexes are assumed to be influenced by SARS-CoV-2 and any metabolic pathways involving these complexes are affected by some degree. At the same time, by applying network propagation methods, including guilt-by-association, network diffusion, and random walks with restart, proteins with high diffusion scores were filtered out in the lung tissue PPI network. The proteins which have close interactions with Krogan proteins are identified as neighbor proteins. Next, Fisher's exact test was performed to derive complexes of interest. Finally, functional enrichment was performed on the complexes of interest and the biological relevance was studied. Additionally, protein-drug databases like DrugBank were used to find drugs by using these complexes as targets.

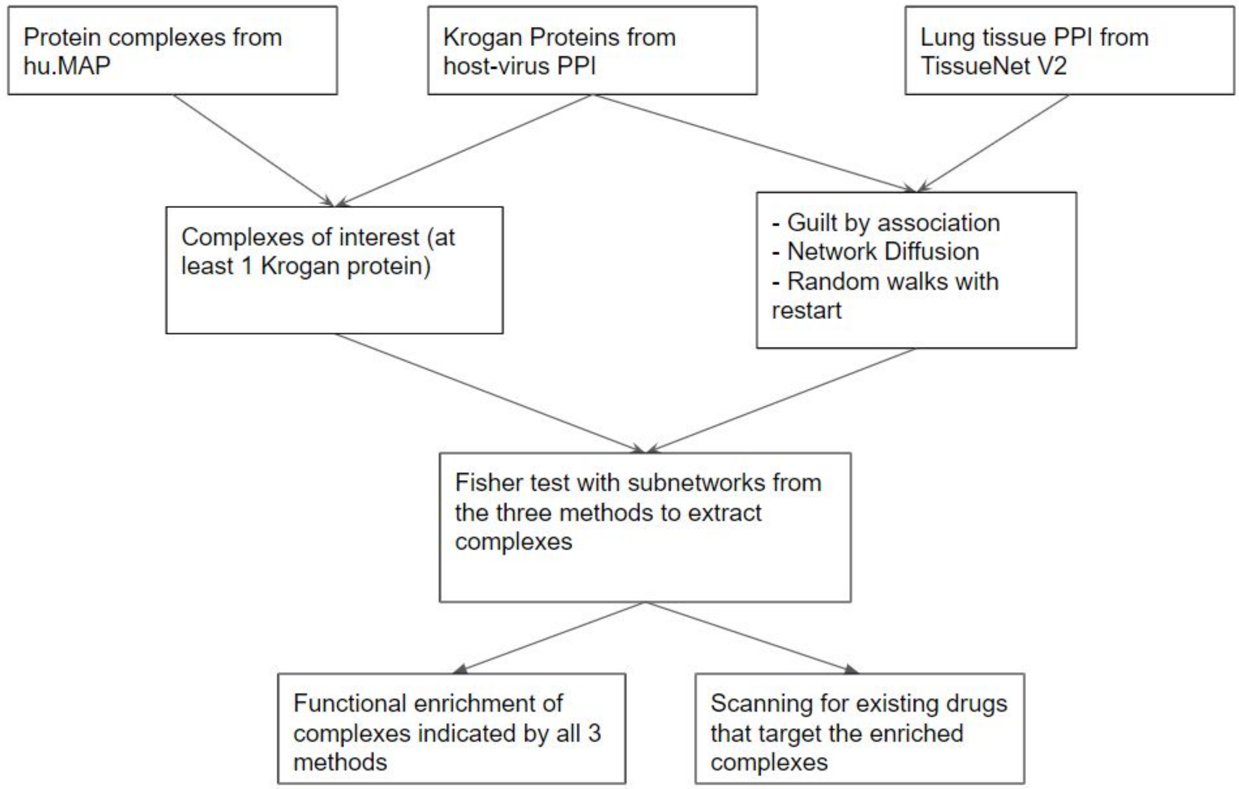

Protein complexes from hu.MAP
Krogan Proteins from host-virus PPI
Lung tissue PPI from TissueNet V2
Complexes of interest (at least 1 Krogan protein)
- Guilt by association
- Network Diffusion
- Random walks with restart
Fisher test with subnetworks from the three methods to extract complexes
Functional enrichment of complexes indicated by all 3 methods
Scanning for existing drugs that target the enriched complexes

Figure 1: Flowchart of methods used to analyze the protein-protein interactions of Krogan proteins

Lung Tissue Protein-Protein Interaction Network

The protein-protein interaction network for the lung tissue was downloaded from the TissueNet v2 database [32]. The database hosts protein-protein interaction networks generated from Human Protein Atlas proteomic data. We used the PPI for lung tissue in our analysis.

Protein Complexes

Protein complexes are groups of proteins that interact with each other to form a multimolecular machine to exert regulatory or structural function [33]. Protein complexes identified as being affected by the virus could be used to study cellular processes being targeted and provide insights into potential existing drugs that are known to work with these complexes. We downloaded all the protein complexes present in the hu.MAP database. These complexes were filtered on the basis of the presence of at least one Krogan protein. We filtered out complexes that include at least one Krogan protein since we have higher confidence that they are affected by SARS-CoV-2.

Network Propagation

By implementing the following three propagation methods, we select roughly the top 5% (around 400) proteins in lung tissue. We assume these proteins are indirectly affected by the SARS-CoV-2 virus through close interaction with multiple Krogan proteins. For the first step, through the lung tissue PPI, a binary adjacency matrix (represented by A) was created to illustrate the interaction of every pair of proteins in lung tissue.

The first approach to identify Krogan-associated proteins is through a guilt-by-association algorithm. Based on the adjacency matrix, each protein interacting with krogan proteins is extracted in lung tissue [25]. Then a histogram was plotted to visualize the distribution of proteins found by guilt-by-association.

Next, network diffusion was implemented to select proteins with the highest diffusion scores. To reduce the complexity of computation, we choose normal diffusion with diffusivity ɑ = 1. The algorithm we use is the same as that covered in the lecture [26].

Adjacency matrix A: a binary matrix representing the interaction between proteins (with an interaction: 1; no interaction: 0)

- Degree matrix D: a diagonal matrix that specifies the number of interactions each protein possess
- Laplacian matrix L = D - A
- Identity matrix I
- Diffusion matrix K: inv(I + 0.1L)

After defining these five matrices, a binary input x (column vector) was created with positions of krogan proteins set to 1 and other positions set to 0. The final diffusion score of all proteins in lung y can be represented as:

$$y = K*x$$

Random Walk with Restart Algorithm

Random walk with restart (RWR), a cutting-edge guilt by association algorithm in network computational biology, is employed to search for novel genes using known genes as seed nodes [13]. In other words, it gives the “closeness” between the two nodes of a graph. As the name suggests, the first step is a “random walk” to the nearest node and there is an option of “restart”, i.e., going back to the starting node. Mathematically, it performs a random walk depending on the value of the “tmax” parameter (maximum number of iterations). The restart parameter expresses the probability of "restarting" from a "core" node at each step of the random walk algorithm. It stops also if the difference of the norm of the probabilities between two consecutive steps is less than “eps”, the maximum allowed difference between the computed probabilities at the steady-state.

Fisher’s Exact Test

Fisher’s Exact Test is a statistical test to determine if two variables have significant associations. The test gives us a low p-value if the association is less likely to occur by chance and a higher p-value for a random association. The test is evaluated using a contingency table. For our analysis, the variables were the protein complex and PPI sub-network indicated by one of the three propagation methods. An example contingency matrix for a complex is given in Table 1. Fisher's test p-value is computed using the scipy.stats.fisher_exact function from the scipy library in python.

Table 1: Contingency table for a protein complex with 19 proteins. This complex has a p-value of 0.01 as given by Fisher's exact test.

	In subnetwork	**Not in subnetwork**	**Totals**
In protein complex	4	15	19
Not in protein complex	466	8610	9076
Totals	470	8625	9095

For each propagation method, we evaluated each protein against the subnetwork and obtained a p-value for each complex. We then found the intersection of the complexes indicated by the three methods to find complexes that were studied as potential drug targets.

RESULTS

From guilt-by-association, 2962 proteins in lung tissue were found to interact with at least one Krogan protein.

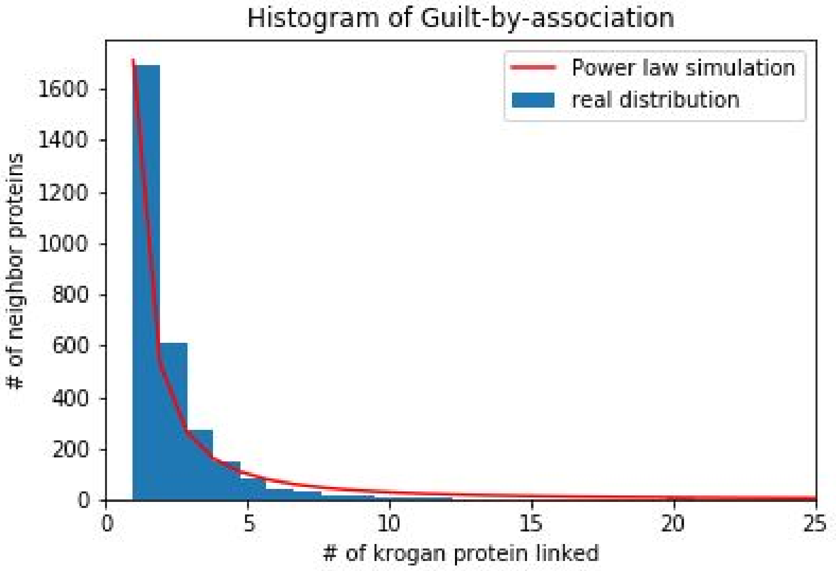


Figure 2. Histogram of Proteins with at least one interaction with Krogan Proteins.

In the histogram shown in figure. 2, x-axis represents the number of Krogan proteins and the y-axis represents the number of lung proteins that the values occurred within the intervals set by the x-axis. The histogram of protein distribution fulfills power-law distribution. This phenomenon indicates the ratio of proteins with 1 more Krogan protein linking to is constant for everywhere in the histogram [27]. The power-law simulation (red line) has the power $\alpha = -1.76$, which has a slightly lower amplitude than normal case ($-2 < \alpha < -3$). Finally, a threshold of 4 was set to select 385 lung proteins that are linked to 4 or more Krogan proteins.

Figure.3 shows part of the guilt-by-association network, which has 14 proteins (blue dots) connecting to > 15 krogan proteins (red dots). Random numbers of other lung proteins (green dots) are also added for each of these 14 proteins. The complete protein-protein interaction diagram of the selected 385 lung proteins and Krogan proteins is provided in the appendix.

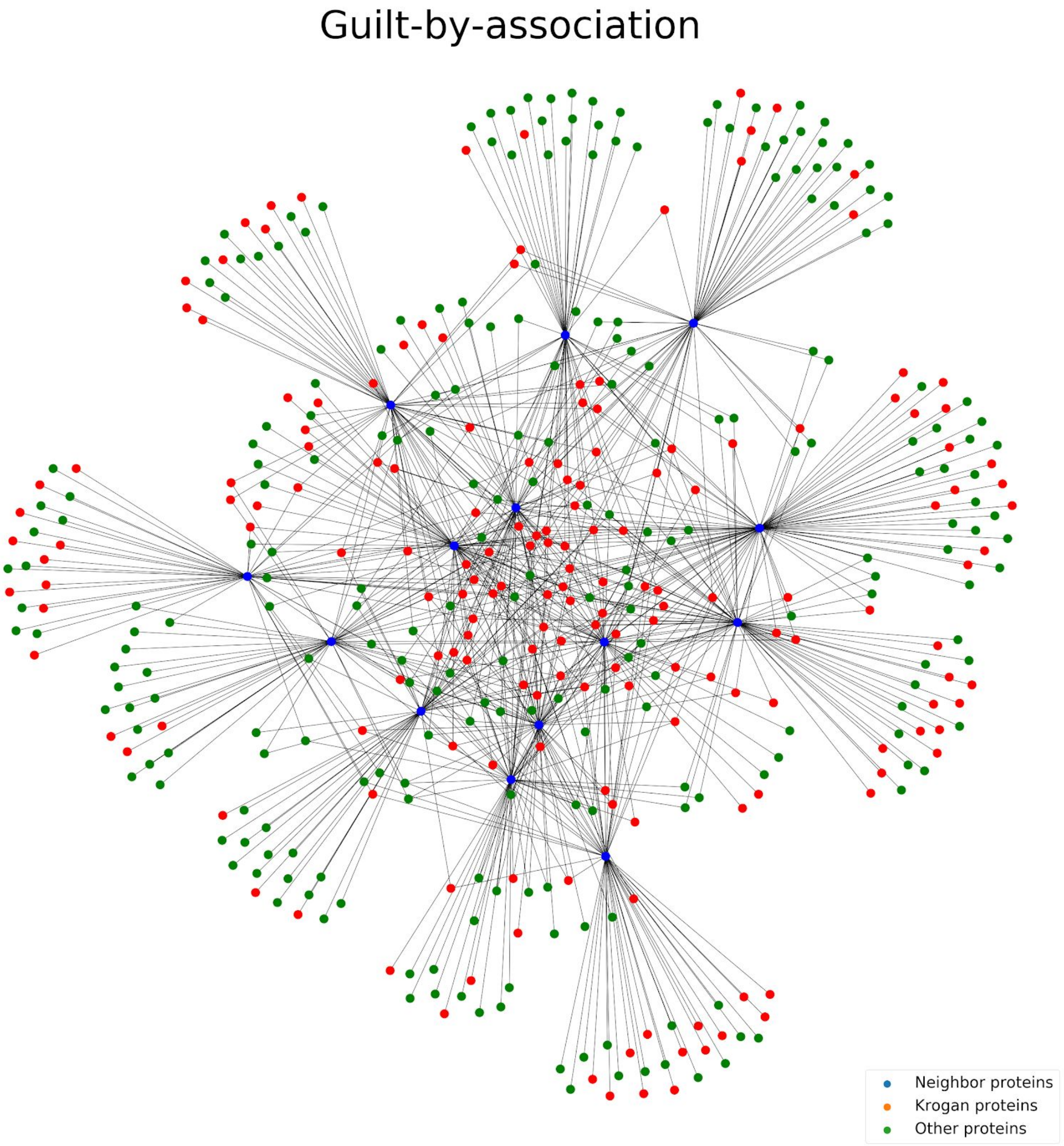


Figure.3 Part of the guilt-by-association network.

The 14 proteins found to interact with Krogan proteins using guilt-by-association were functionally enriched using DAVID. Five clusters of pathways were identified to be significantly related to these proteins. The highest enriched cluster had pathway interactions with acetylation and the nucleus. The proteins were also found to interact with cytoskeletal organization of microtubules and centrosomes, as well as regulation of mitotic cell division. Ubiquitin protein activity pathways were also associated, as well as Dwarfism, disease mutation and polymorphism. These pathways are common in the epithelium and lymphocytes, both first lines of defense in the immune response to SARS-CoV-2 virus.

For network diffusion, a threshold of 0.05 was chosen to select 335 neighbor proteins. Figure.4 illustrates the sample network diffusion diagram. The final diffusion score is represented by the color of nodes: a darker node color correlates to a higher score, which means the corresponding protein has a closer interaction with krogan protein. Names of 18 proteins with the highest scores are labeled.

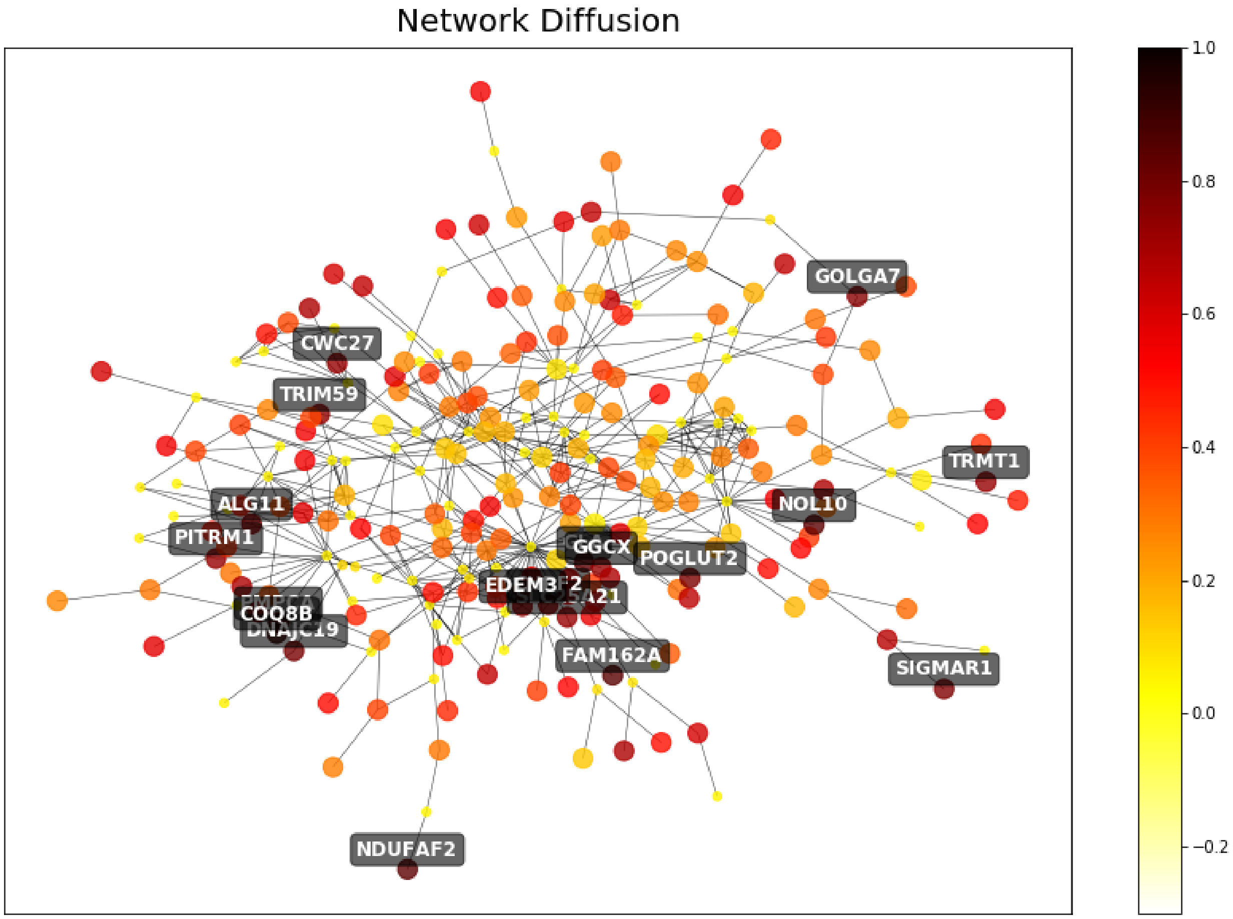


Figure 4: Subnetwork obtained by network diffusion with diffusion score threshold 0.005. Larger nodes are Krogan proteins and smaller nodes are other proteins present in the PPI. The color of a node is based on the diffusion score with a higher score being closer to red and a lower score closer to yellow. 18 proteins with diffusion scores greater than 0.7 are labeled.

The 18 most highly enriched protein complexes found from the network diffusion algorithm were then functionally enriched for biological analysis. Three pathway clusters were highly enriched. There were significant relationships between the protein complexes and mitochondrial transit peptides. The proteins were also related to the membrane proteins and the transmembrane region. They were also related to glycoproteins in the membrane. The pathways that interact with the proteins from the network diffusion were related to lymphoblasts and mammary glands.

Finally, the random walk with restart (RWR) was employed to determine the protein complexes based on their proximity to the Krogan proteins. RWR is another guilt-by-association method which is used to search for novel genes using known genes as seed nodes. The adjacency matrix was created using the krogan proteins and the lung PPI obtained from the HPA proteome database. This was then used to run the RWR algorithm. The threshold was set to 0.0005 and 435 proteins were found.

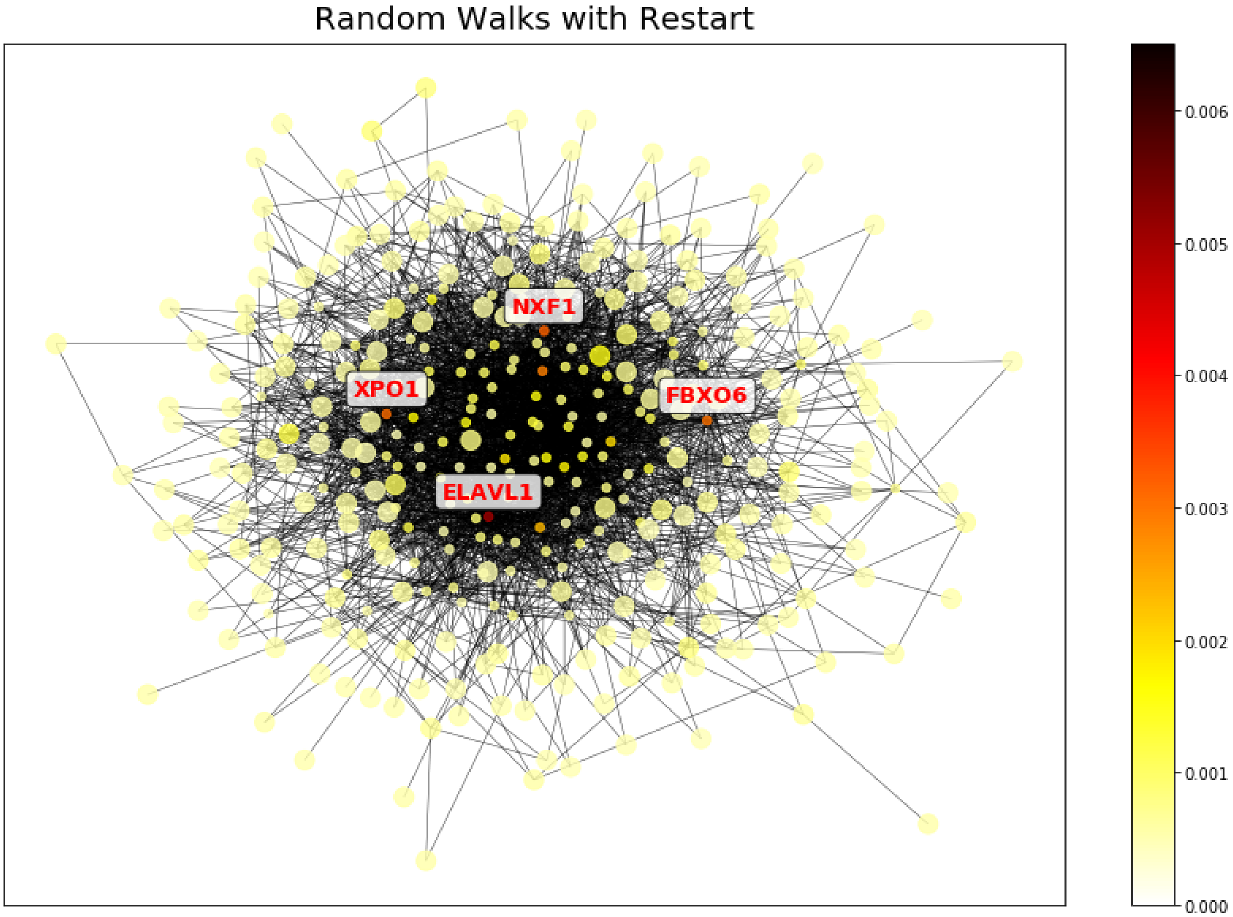


Figure 5: Subnetwork obtained by applying a threshold of 0.0005 on the random walk with restarts score. Larger nodes are Krogan proteins and smaller nodes are other proteins present in the PPI. The color of a node is based on the RWR score with a higher score being closer to red and a lower score closer to yellow. 4 proteins with RWR score greater than 0.004 are labeled.

The top four significant proteins obtained from running the RWR algorithm, depicted in the graph above, are discussed below.

NXF1 is a major component of the nuclear mRNA export pathway that is taken over by many viruses whose life cycles include nuclear stages [6]. NXF1 is necessary for viral protein expression, but not for viral RNA synthesis which might explain its significance in COVID-19.

Xpo1 or exportin1 is a key protein involved in nucleocytoplasmic transport. Attempts have been made to exploit this protein since it can be useful in quarantining the key viral accessory proteins and genomic materials in the nucleus of the host cell thus reducing virus replication and immuno-pathogenicity [7].

FBXO6 encodes a member of the F-box protein family which is responsible for the endoplasmic reticulum-associated degradation pathway (ERAD) for misfolded lumenal proteins. ER stress occurs in the event of viral infection which explains the significance of this gene [8].

ELAVL1 codes for ELAV-like RNA binding protein 1 that is known to regulate T cell activation via the NOD-like receptor (NLR) [9].

Protein complexes of interest

Protein complexes were filtered by the presence of at least one krogan protein. After filtering 740 complexes were obtained that had at least one krogan protein present in them.

Fisher Exact Test to find relevant complexes

The Fisher exact test is computed to assign a p-value to each complex. We picked the complexes with a p-value < 0.05 as enriched in the network. This is similar to the ideas exhibited in Liu et al [5]. In the guilt-by-association subnetwork, 33 complexes were found to be enriched. In the network diffusion and RWR subnetworks, 132 and 128 complexes were enriched respectively. Out of these 12 complexes were found to be common among all three networks. We further performed functional enrichment on these complexes.

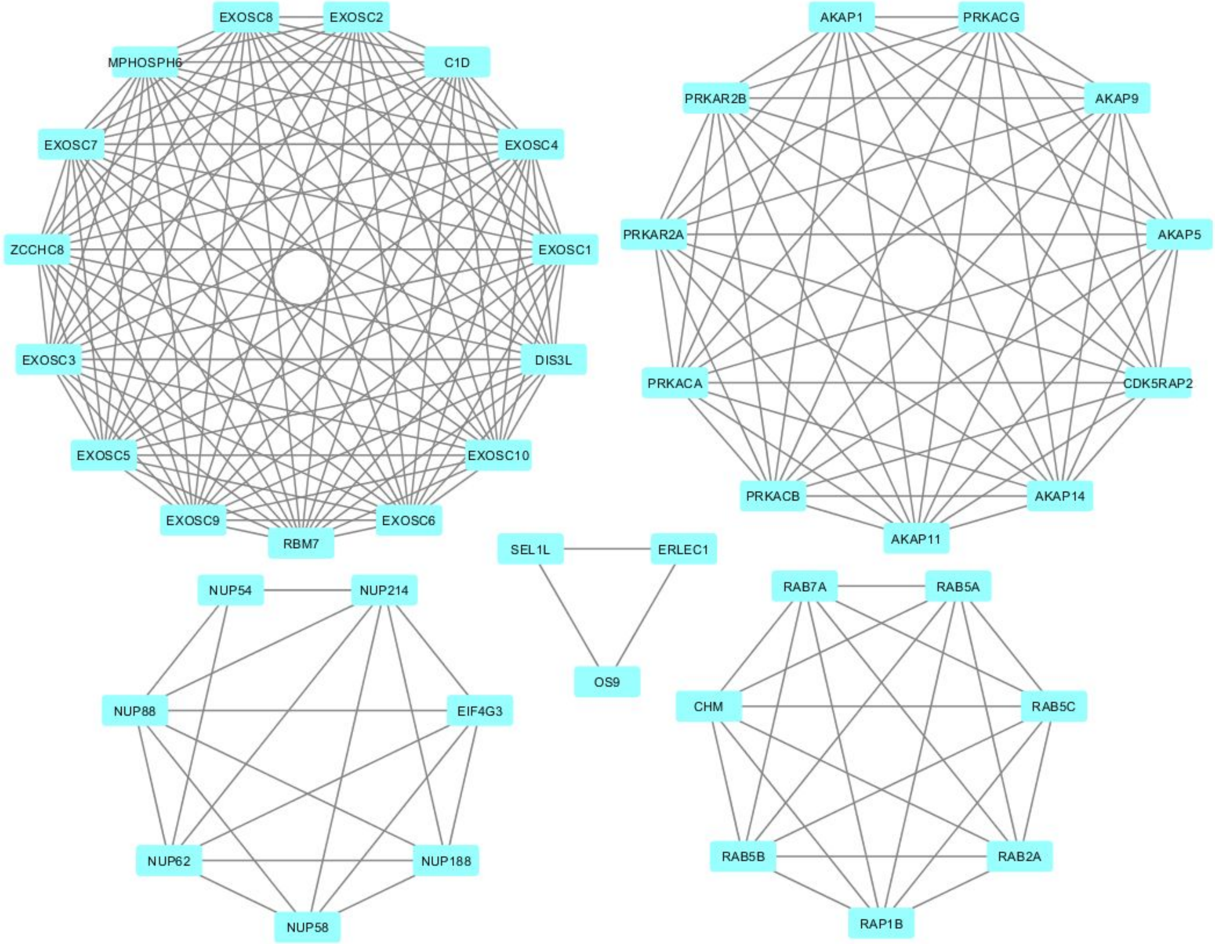


Figure 6: The 12 complexes represented as a single network. Since some of the complexes had common proteins, we finally arrived at 5 disconnected complexes as shown.

Functional Enrichment

In order to determine the function and areas of interest the complexes were involved in, we used DAVID [38, 39] and Cytoscape [40] to perform functional annotation and visualization of related pathways respectively. Of the 12 complexes determined to be enriched by the guilt-by-association subnetwork, the RWR subnetwork, and the network diffusion subnetwork, some protein complexes were repeated. This resulted in 5 distinct complexes that we analyzed for function and drug targeting, as seen in Figure 6.

In order to determine lung tissues potentially affected by SARS-CoV-2 based on the protein complex networks, we used DAVID, a functional annotation tool [38, 39], to retrieve linear chart reports of selected annotation categories of tissue expression. We then visualized the affected tissues using Cytoscape [40] with EnrichmentMap [41] to create a radial heatmap weighted by p-value with a cutoff value of 0.001. As seen in Figure 7, the potentially affected tissues by the 5 enriched protein complexes include the brain, epithelium, lung, placenta, testes, and sperm.

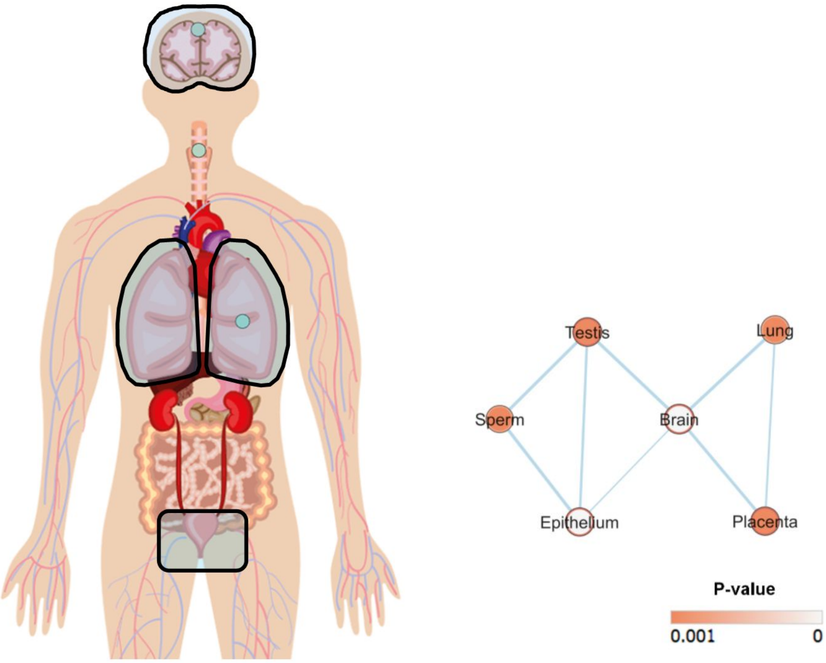


Figure 7: The tissue areas enriched by all 5 protein complexes determined, shown by the highlighted areas

The pathways affected by SARS-CoV-2 are of interest as well, as possible treatment methods can act by targeting genetic pathways to inhibit or express genes that may contribute to symptoms.

Each of the 5 complexes were analyzed for enriched pathways using DAVID [38, 39], then the results were visualized in Cytoscape [40] with EnrichmentMap [41]. The orange nodes have the highest p values and are the closest to the significance cutoff of 0.001.

Complex I: AKAP9, PRKACG, AKAP1, PRKAR2B, PRKAR2A, PRKACA, PRKACB, AKAP11, AKAP14, CDK5RAP2, and AKAP5.

This complex is involved in PKA signaling pathways. It is also involved in taste transduction and salivary secretion. This is interesting since the loss of taste and smell is a unique symptom to the COVID-19. The complex is involved in hormone regulation, such as insulin secretion, progesterone, aldosterone, and thyroid hormone synthesis. This complex also affects renin secretion and vasopressin-regulated water reabsorption [38, 39]. This network of related pathways can be seen in Figure 8.

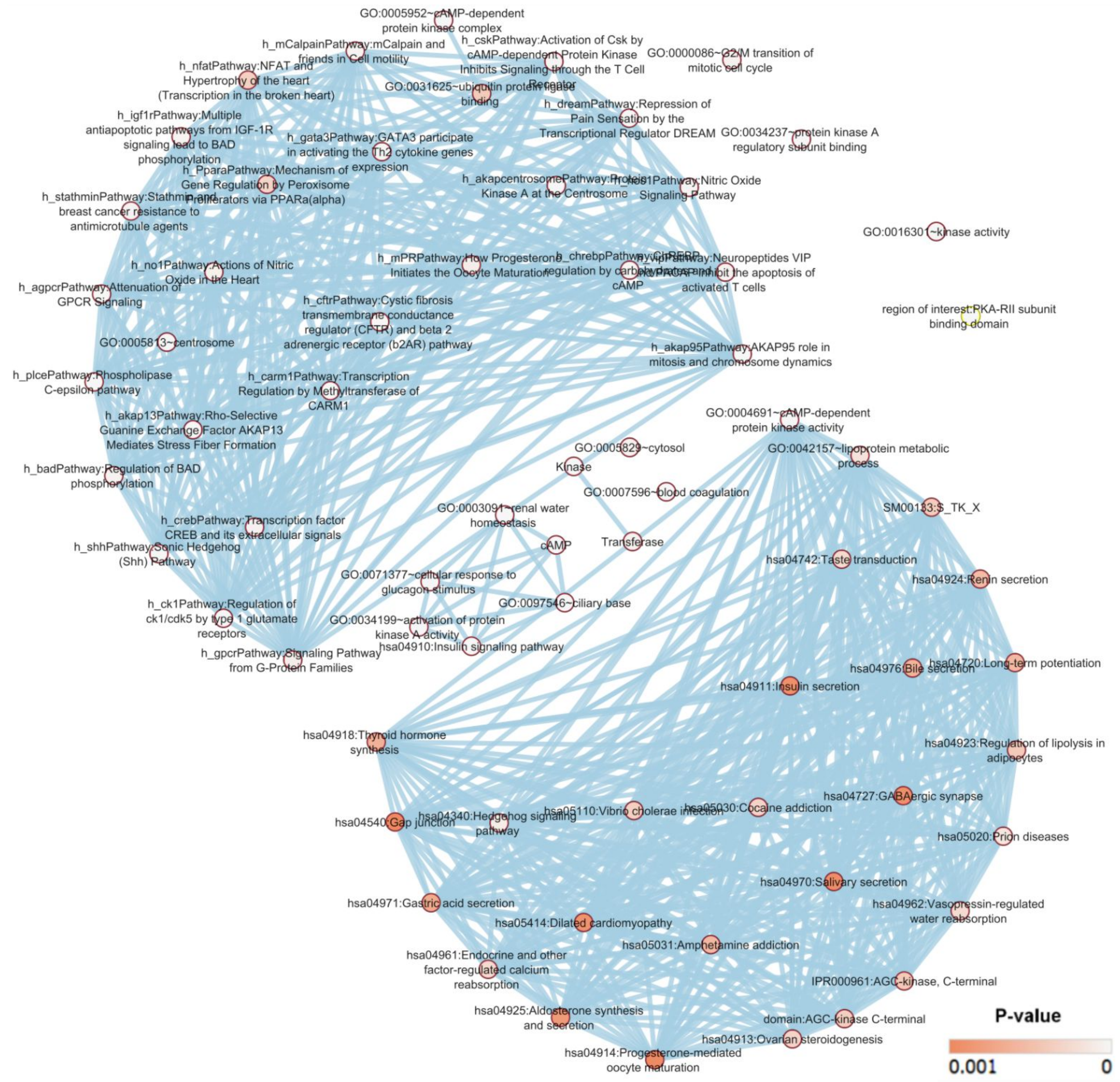

GO:0005952~cAMP-dependent protein kinase complex
h_mCalpainPathway:mCalpain and friends in Cell motility
h_cskPathway:Activation of Csk by cAMP-dependent Protein Kinase Inhibits Signaling through the T Cell Receptor
GO:0000086~G2/M transition of mitotic cell cycle
h_nfatPathway:NFAT and Hypertrophy of the heart (Transcription in the broken heart)
GO:0031625~ubiquitin protein ligase binding
h_dreamPathway:Repression of Pain Sensation by the Transcriptional Regulator DREAM
h_igf1rPathway:Multiple antiapoptotic pathways from IGF-1R signaling lead to BAD phosphorylation
GO:0034237~protein kinase A regulatory subunit binding
h_gata3Pathway:GATA3 participate in activating the Th2 cytokine genes expression
h_PparaPathway:Mechanism of Gene Regulation by Peroxisome Proliferators via PPARa(alpha)
h_akapcentrosomePathway:Protein Kinase A at the Centrosome
h_nos1Pathway:Nitric Oxide Signaling Pathway
h_stathminPathway:Stathmin and breast cancer resistance to antimicrotubule agents
GO:0016301~kinase activity
h_no1Pathway:Actions of Nitric Oxide in the Heart
h_mPRPathway:How Progesterone Initiates the Oocyte Maturation
h_chrebpPathway:ChREBP regulation by carbohydrates and cAMP
h_vipPathway:Neuropeptides VIP and PACAP inhibit the apoptosis of activated T cells
h_agpcrPathway:Attenuation of GPCR Signaling
region of interest:PKA-RII subunit binding domain
h_cftrPathway:Cystic fibrosis transmembrane conductance regulator (CFTR) and beta 2 adrenergic receptor (b2AR) pathway
GO:0005813~centrosome
h_akap95Pathway:AKAP95 role in mitosis and chromosome dynamics
h_plcePathway:Phospholipase C-epsilon pathway
h_carm1Pathway:Transcription Regulation by Methyltransferase of CARM1
GO:0004691~cAMP-dependent protein kinase activity
h_akap13Pathway:Rho-Selective Guanine Exchange Factor AKAP13 Mediates Stress Fiber Formation
GO:0005829~cytosol
GO:0042157~lipoprotein metabolic process
Kinase
h_badPathway:Regulation of BAD phosphorylation
GO:0007596~blood coagulation
GO:0003091~renal water homeostasis
SM00133:S_TK_X
h_crebPathway:Transcription factor CREB and its extracellular signals
cAMP
Transferase
h_shhPathway:Sonic Hedgehog (Shh) Pathway
hsa04742:Taste transduction
hsa04924:Renin secretion
GO:0071377~cellular response to glucagon stimulus
h_ck1Pathway:Regulation of ck1/cdk5 by type 1 glutamate receptors
GO:0097546~ciliary base
GO:0034199~activation of protein kinase A activity
hsa04910:Insulin signaling pathway
h_gpcrPathway:Signaling Pathway from G-Protein Families
hsa04976:Bile secretion
hsa04720:Long-term potentiation
hsa04911:Insulin secretion
hsa04918:Thyroid hormone synthesis
hsa04923:Regulation of lipolysis in adipocytes
hsa04727:GABAergic synapse
hsa04540:Gap junction
hsa04340:Hedgehog signaling pathway
hsa05110:Vibrio cholerae infection
hsa05030:Cocaine addiction
hsa05020:Prion diseases
hsa04970:Salivary secretion
hsa04971:Gastric acid secretion
hsa04962:Vasopressin-regulated water reabsorption
hsa05414:Dilated cardiomyopathy
hsa05031:Amphetamine addiction
hsa04961:Endocrine and other factor-regulated calcium reabsorption
IPR000961:AGC-kinase, C-terminal
hsa04925:Aldosterone synthesis and secretion
domain:AGC-kinase C-terminal
hsa04913:Ovarian steroidogenesis
hsa04914:Progesterone-mediated oocyte maturation
P-value
0.001
0

Figure 8: The functional pathways found in the first complex (AKAP9 PRKACG AKAP1 PRKAR2B PRKAR2A PRKACA PRKACB AKAP11 AKAP14 CDK5RAP2 AKAP5). This complex primarily consists of PKA signaling pathways.

Complex II: CHM, RAB5B, RAB5A, RAB5C, RAP1B, RAB7A, and RAB2A.

This complex is involved in metabolic signaling and cellular transport pathways. It is significantly related to GTP/GDP binding activities for intracellular signalling. It is also related to endocytotic vesicles, phagosomes, melanosomes, and endosomes. The complex is related to Ras signaling, amoebiasis and lipoprotein regulation [38, 39]. This network of related pathways can be seen in Figure 9.

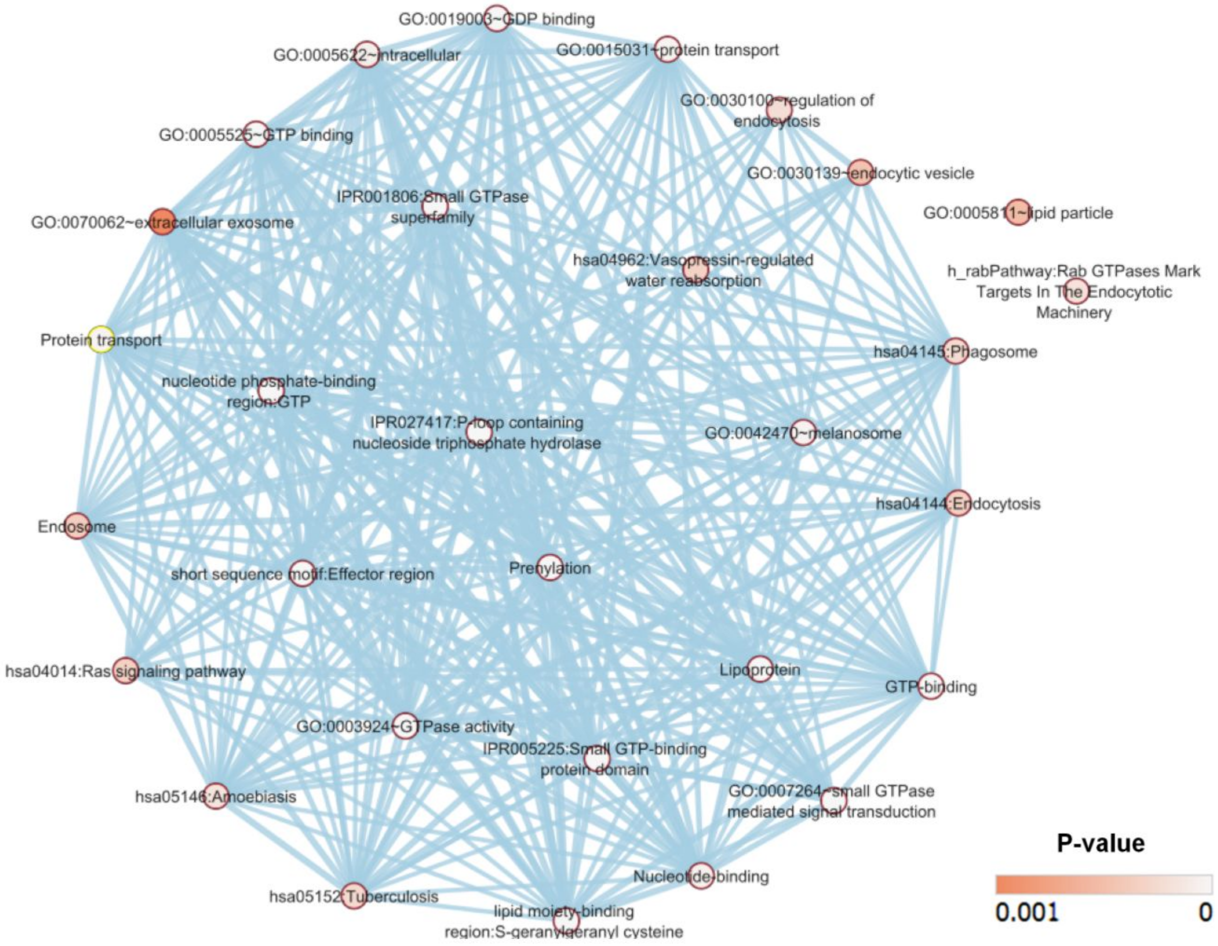

GO:0019003~GDP binding
GO:0005622~intracellular
GO:0015031~protein transport
GO:0030100~regulation of endocytosis
GO:0005525~GTP binding
GO:0030139~endocytic vesicle
IPR001806:Small GTPase superfamily
GO:0070062~extracellular exosome
GO:0005811~lipid particle
hsa04962:Vasopressin-regulated water reabsorption
h_rabPathway:Rab GTPases Mark Targets In The Endocytotic Machinery
Protein transport
hsa04145:Phagosome
nucleotide phosphate-binding region:GTP
IPR027417:P-loop containing nucleoside triphosphate hydrolase
GO:0042470~melanosome
hsa04144:Endocytosis
Endosome
short sequence motif:Effector region
Prenylation
hsa04014:Ras signaling pathway
Lipoprotein
GTP-binding
GO:0003924~GTPase activity
IPR005225:Small GTP-binding protein domain
GO:0007264~small GTPase mediated signal transduction
hsa05146:Amoebiasis
Nucleotide-binding
hsa05152:Tuberculosis
lipid moiety-binding region:S-geranylgeranyl cysteine
P-value
0.001
0

Figure 9: The functional pathways found in the second complex (CHM RAB5B RAB5A RAB5C RAP1B RAB7A RAB2A). This complex primarily consists of metabolic signaling and transport pathways.

Complex III: ERLEC1, SEL1L, and OS9. This complex is involved in endoplasmic reticulum quality control, ER associated catabolism, and ER protein processing in the ER. This complex has significant relations to the ubiquitin ligase complex, phosphate binding domains, and the glucosidase II beta subunit [38, 39]. This network of related pathways can be seen in Figure 10.

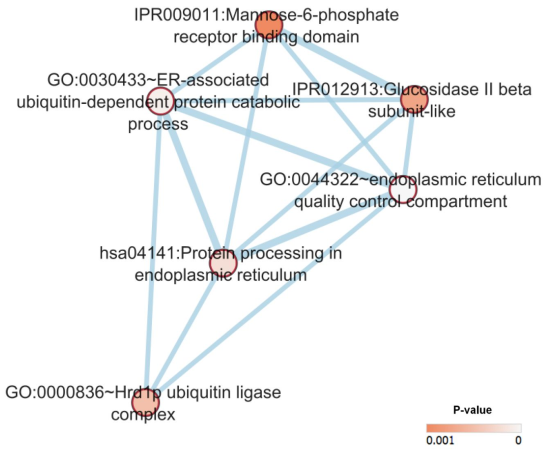

IPR009011:Mannose-6-phosphate receptor binding domain
GO:0030433~ER-associated ubiquitin-dependent protein catabolic process
IPR012913:Glucosidase II beta subunit-like
GO:0044322~endoplasmic reticulum quality control compartment
hsa04141:Protein processing in endoplasmic reticulum
GO:0000836~Hrd1p ubiquitin ligase complex
P-value
0.001
0

Figure 10: The functional pathways found in the third complex (ERLEC1 SEL1L OS9). This complex primarily consists of endoplasmic reticulum (ER) pathways.

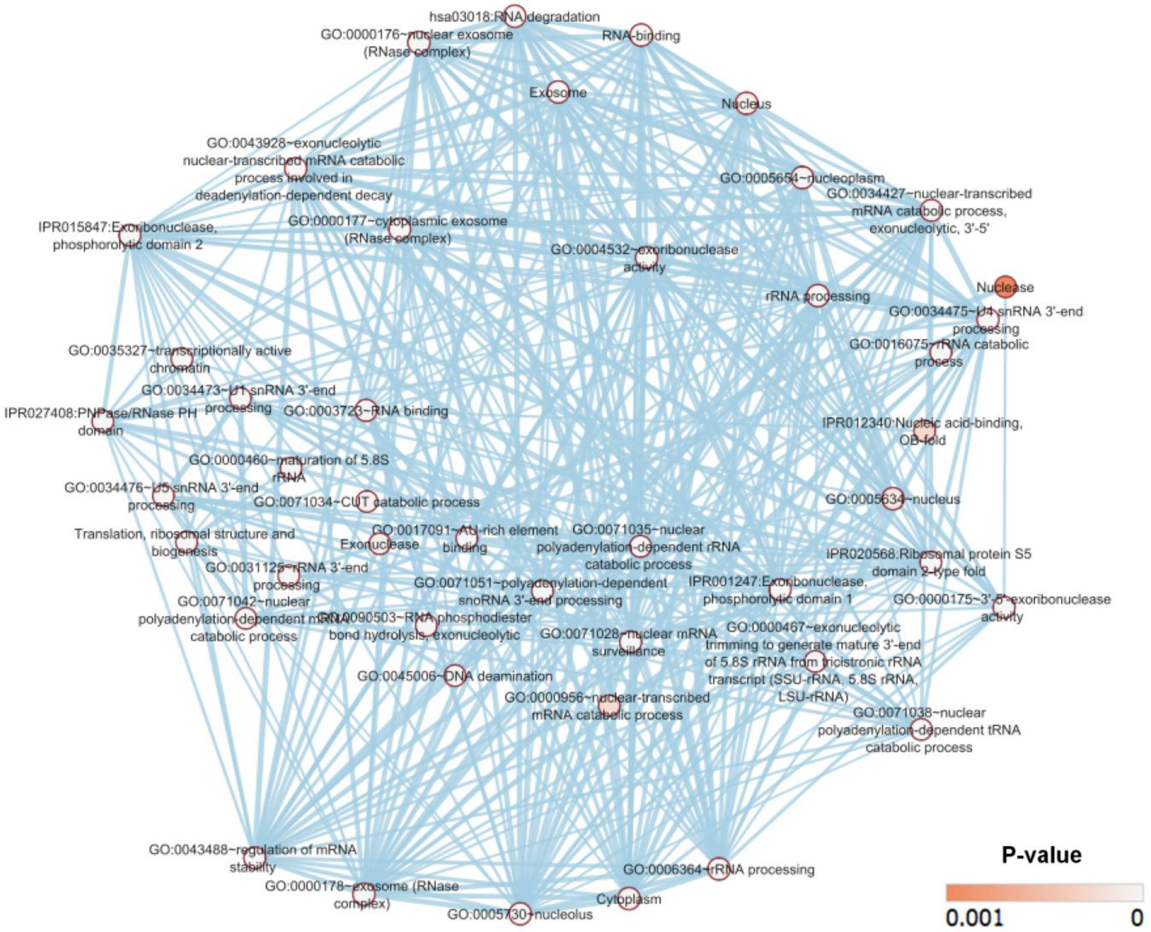

hsa03018:RNA degradation
GO:0000176~nuclear exosome (RNase complex)
RNA-binding
Exosome
Nucleus
GO:0043928~exonucleolytic nuclear-transcribed mRNA catabolic process involved in deadenylation-dependent decay
GO:0005654~nucleoplasm
GO:0034427~nuclear-transcribed mRNA catabolic process, exonucleolytic, 3'-5'
IPR015847:Exoribonuclease, phosphorolytic domain 2
GO:0000177~cytoplasmic exosome (RNase complex)
GO:0004532~exoribonuclease activity
Nuclease
rRNA processing
GO:0034475~U4 snRNA 3'-end processing
GO:0016075~rRNA catabolic process
GO:0035327~transcriptionally active chromatin
GO:0034473~U1 snRNA 3'-end processing
IPR027408:PNPase/RNase PH domain
GO:0003723~RNA binding
IPR012340:Nucleic acid-binding, OB-fold
GO:0000460~maturation of 5.8S rRNA
GO:0034476~U5 snRNA 3'-end processing
GO:0071034~CUT catabolic process
GO:0005634~nucleus
Translation, ribosomal structure and biogenesis
GO:0017091~AU-rich element binding
Exonuclease
GO:0071035~nuclear polyadenylation-dependent rRNA catabolic process
IPR020568:Ribosomal protein S5 domain 2-type fold
GO:0031125~rRNA 3'-end processing
GO:0071051~polyadenylation-dependent snoRNA 3'-end processing
IPR001247:Exoribonuclease, phosphorolytic domain 1
GO:0000175~3'-5'-exoribonuclease activity
GO:0071042~nuclear polyadenylation-dependent mRNA catabolic process
GO:0090503~RNA phosphodiester bond hydrolysis, exonucleolytic
GO:0071028~nuclear mRNA surveillance
GO:0000467~exonucleolytic trimming to generate mature 3'-end of 5.8S rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)
GO:0045006~DNA deamination
GO:0000956~nuclear-transcribed mRNA catabolic process
GO:0071038~nuclear polyadenylation-dependent tRNA catabolic process
GO:0043488~regulation of mRNA stability
GO:0006364~rRNA processing
GO:0000178~exosome (RNase complex)
Cytoplasm
GO:0005730~nucleolus
P-value
0.001
0

Figure 11: The functional pathways found in the fourth complex (EXOSC9 RBM7 EXOSC6 EXOSC10 EXOSC3 DIS3L EXOSC1 EXOSC4 C1D EXOSC2 EXOSC8 EXOSC5 MPHOSPH6 EXOSC7 ZCCHC8). This complex primarily consists of RNA exosome machinery pathways.

Complex IV: EXOSC9, RBM7, EXOSC6, EXOSC10, EXOSC3, DIS3L, EXOSC1, EXOSC4, C1D, EXOSC2, EXOSC8, EXOSC5, MPHOSPH6, EXOSC7, and ZCCHC8.

Exosome machinery in the cytoplasm, nucleus, and nucleolus were highly related, as well many ribosomal catabolic and degradation processes [38, 39]. This network of related pathways can be seen above in Figure 11.

COMPLEX V: EIF4G3, NUPL1, NUP54, NUP68, NUP88, NUP188, and NUP214.

This complex is involved in nuclear pore complexes. It has significant relations to transport into the nucleus, in many pathways including tRNA, protein, glucose transport. It is also involved in gene silencing and mitotic nuclear envelope disassembly. It has interesting relationships with viral transcription and intracellular transport of viruses. It also contains a link to the regulation of cellular responses to heat. This may be related to the immune response to the SARS-CoV-2 virus of fevers [38, 39]. This network of related pathways can be seen in Figure 12.

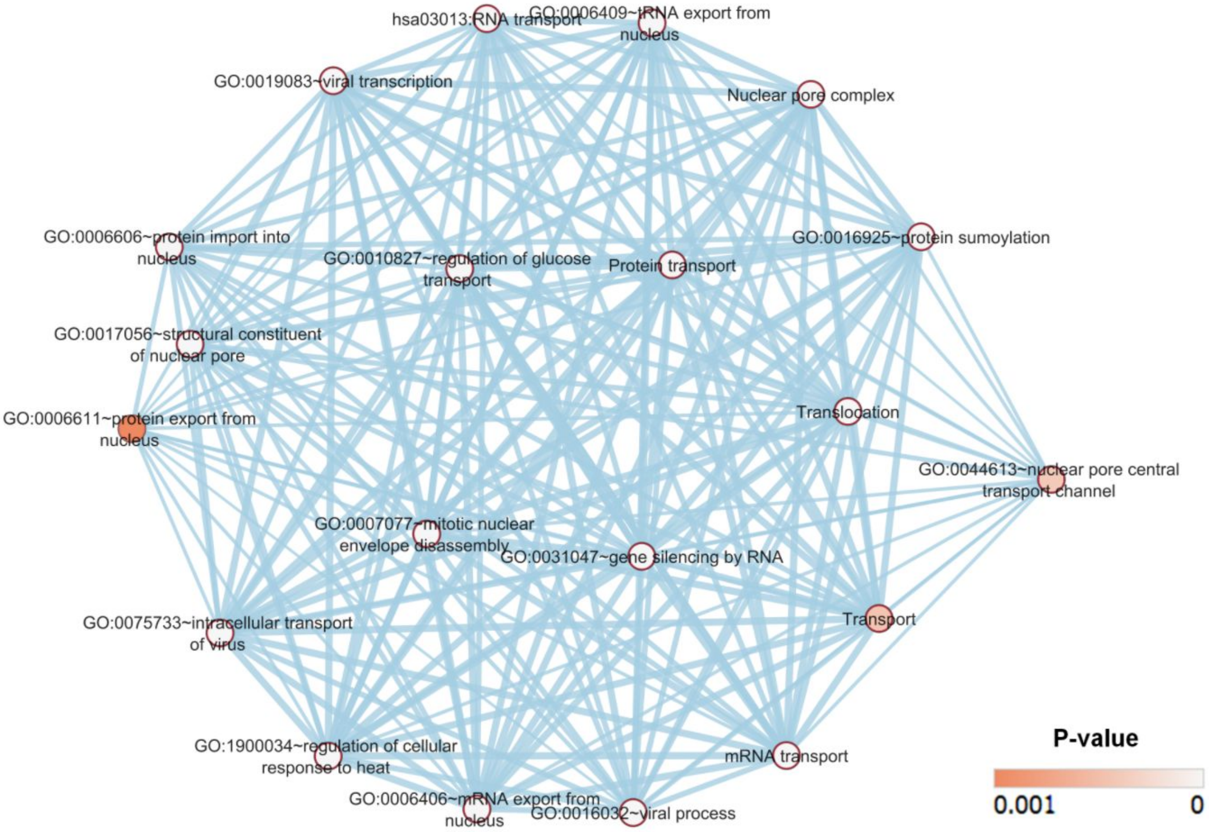

hsa03013:RNA transport
GO:0006409~tRNA export from nucleus
GO:0019083~viral transcription
Nuclear pore complex
GO:0006606~protein import into nucleus
GO:0010827~regulation of glucose transport
Protein transport
GO:0016925~protein sumoylation
GO:0017056~structural constituent of nuclear pore
GO:0006611~protein export from nucleus
Translocation
GO:0044613~nuclear pore central transport channel
GO:0007077~mitotic nuclear envelope disassembly
GO:0031047~gene silencing by RNA
GO:0075733~intracellular transport of virus
Transport
GO:1900034~regulation of cellular response to heat
mRNA transport
GO:0006406~mRNA export from nucleus
GO:0016032~viral process
P-value
0.001
0

Figure 12: The functional pathways found in the fifth complex (EIF4G3 NUPL1 NUP54 NUP68 NUP88 NUP188 NUP214). This complex primarily consists of nuclear pore complexes.

Protein Target- Drug Analysis

With the growing number of cases, drugs might prove to be useful but their efficacy must be studied thoroughly through randomized clinical trials. Following this, it is imperative that the drug be provided to patients belonging to the stage of the disease for which its efficacy was demonstrated. One such example is that of Dexamethasone wherein it was found to reduce mortality in severely ill patients but was later seen to increase the mortality in case of mild to moderately sick non-hospitalized patients [24].

A recent research showed multiple stages of the virus' replication cycle at which various drugs can effectively suppress its growth and subsequent spread.

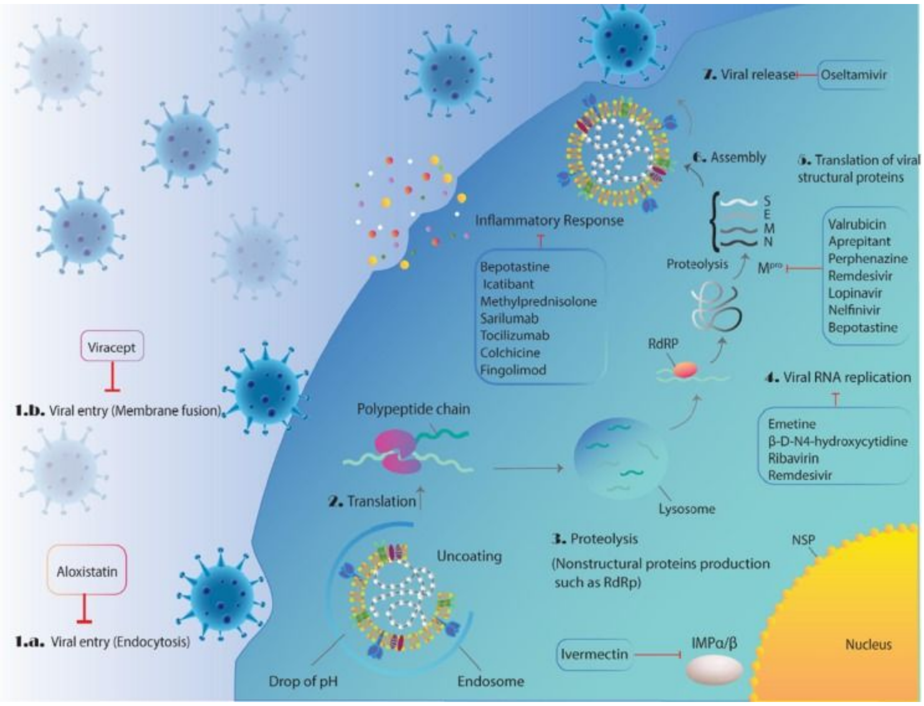
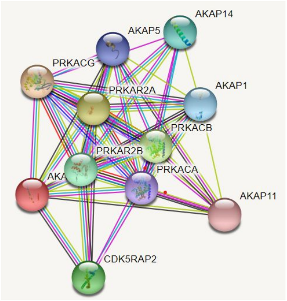
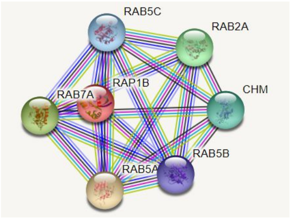
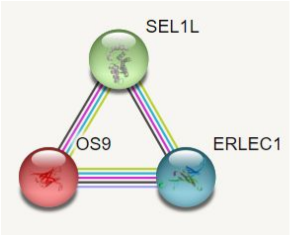
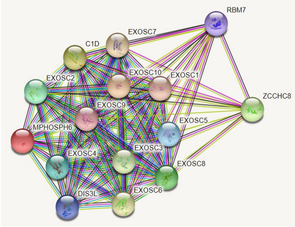
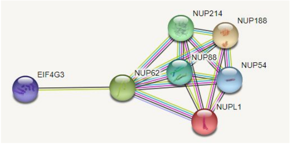


Figure : Schematic overview of the potential therapeutic drugs undergoing clinical trial or have been proposed against SARS-CoV-2 virus with their mechanism of action in different main steps of CoV life cycle in host cells [24]

The following table includes the protein complexes obtained after running all three analyses, the pathways they are a part of, their functions, and the existing drugs that can be used keeping these complexes as targets. STRING, a database consisting of known and predicted protein-protein interactions. These interactions can be either direct, i.e. physical, or indirect, i.e. functional [10].

Table 1: Protein complexes filtered after running the three algorithms, its representation in a network using STRING, the pathway it is a part of, its functions, and the existing drugs that can be used for these complexes as targets.

Protein Complex	Protein Complex Network (STRING)	Pathway	Function	Existing Drugs
AKAP9 PRKACG AKAP1 PRKAR2B PRKAR2A PRKACA PRKACB AKAP11 AKAP14 CDK5RAP2 AKAP5	 AKAP14 AKAP5 PRKACG PRKAR2A AKAP1 PRKACB PRKAR2B PRKACA AKA AKAP11 CDK5RAP2	Protein kinase A (PKA) signaling pathway	Regulation of cell cycle and proliferation, metabolism, transmission of nerve impulses, cytoskeleton remodeling, muscle contraction, cell survival, and other cell processes. [11]	Acetaminophen- Provides relief in the case of muscle pain, inflammation, fever, chills, cough/cold, respiratory tract infection [12].
CHM RAB5B RAB5A RAB5C RAP1B RAB7A RAB2A		Metabolic signaling and vesicular transport	Key regulators of eukaryotic membrane trafficking -budding, transport, docking, and fusion of lipid bilayer vesicles [13]	Ticagrelor- Reduces sepsis-induced lung injury, and neutrophils pulmonary infiltration [14] Vidarabine- nucleoside antibiotic that blocks DNA polymerase and also acts as a chain terminator in DNA replication [22]
ERLEC1 SEL1L OS9		Endoplasmic reticulum (ER) quality control and ER-associated degradation (ERAD)	Binds terminally misfolded non-glycosylated proteins as well as improperly folded glycoproteins, retain them in the ER, and possibly	Selenium and Selenoprotein- Antioxidant, anti-inflammatory effects, augments immune cell function [16]

			transfer them to the ubiquitination machinery and promote their degradation [15]	
EXOSC9 RBM7 EXOSC6 EXOSC10 EXOSC3 DIS3L EXOSC1 EXOSC4 C1D EXOSC2 EXOSC8 EXOSC5 MPHOSPH6 EXOSC7 ZCCHC8	 RBM7 EXOSC7 C1D EXOSC10 EXOSC1 EXOSC2 ZCCHC8 EXOSC9 EXOSC5 MPHOSPH6 EXOSC3 EXOSC4 EXOSC8 DIS3L EXOSC6	RNA exosome complex-RNA degradation machinery	RNA quality control in the nucleus. It also degrades many types of cryptic transcripts that are generated as a result of pervasive transcription and remove aberrant RNA molecules that failed to mature properly[17]	Salmeterol-treatment of asthma and chronic obstructive pulmonary disease COPD [18]
EIF4G3 NUPL1 NUP54 NUP68 NUP88 NUP188 NUP214	 NUP214 NUP188 NUP88 EIF4G3 NUP62 NUP54 NUPL1	Nuclear pore complexes	Nuclear transport, differentiation, gene expression, chromatin organization, epigenetic regulation, replication-coupled DNA repair, and mitosis [19]	Sodium Fluoride; Chlorhexidine - antiseptic mouthwash or gargling reduced oral viral load [20]

DISCUSSION

SARS-CoV-2 is a multi-systemic infection that developed into a pandemic sooner than predicted.

Studies providing epidemiological, virological and pathophysiological characteristics of the virus is the need of the hour. Our work applied network propagation algorithms including guilt-by-association, network diffusion, and random walk with restart in order to describe the interactome of coronavirus using Krogan proteins, a predefined protein dataset known to be affected by this virus. Ascertaining the role of genes and proteins in our body is required for understanding the molecular and biochemical processes that maintain homeostasis or cause diseases, identifying drug targets, and coming up with reliable therapeutic agents to reverse the imbalance. Protein complexes play important roles in numerous biological processes like protein synthesis, signaling and cellular degradation processes. The analyses of complexes in tandem with the networks using Krogan proteins in the lung tissue PPI helped us understand the tissues and pathways being affected by the SARS-CoV-2 virus.

The major organs affected by the complexes determined by the original lung tissue data are illustrated in Fig 7. The lungs are the most expected target of the virus since it is known to be a lower respiratory tract infection. It makes sense that the lung tissue protein complexes with interactions with Krogan proteins, would have a feedback loop of self interaction. The brain was significantly affected by complexes in the lungs associated with Krogan proteins. The SARS-CoV-2 interaction with the whole brain is not annotated in literature as well as the lungs [37], and more research is needed to determine how the virus affects the brain. Epithelial tissue was also significantly affected. Epithelial tissue exists in the lining of all organs in the body. Specifically vulnerable would be the tissue in the mouth, nose, and respiratory tract. SOme research suggests the virus directly infects the olfactory sensory neurons in the olfactory bulb [42]. Reproductive organs were also significantly affected in the enrichment analysis of affected proteins, including the testes, sperm, and placenta. This suggests possible complications in the male reproductive organs and in pregnant females [36].

The three methods for network propagation used in the project have been used before to select virus-related proteins and expand PPI networks. As for the selected proteins by these methods, most of them are highly associated with RNA export, protein transport, protein degradation and cell-cycle regulation [28]. These pathways are related to SARS-CoV-2 virus targets in lung tissue [29]. Therefore, there is high confidence that the proteins used in fisher exact test, functional enrichment, and drug analysis is affected by the virus.

While safe and effective therapeutics to target SARS-CoV-2 are still being developed, using the existing drugs is preferred since these drugs are already in the market with well established short and long term clinical effects along with risks and contraindications. Also, existing drugs are more affordable costs for public health systems. Moreover, clinicians are more likely to be familiarized with them, making them feasible and more desirable for drug therapy [23]. There are more than 400 drugs and 30 biological agents that are in clinical trials, developed using the principles used to make antiviral drugs in the past. However, they lack a long term safety profile and their affordability which will be a major hindrance in its massive public usage. In our study the protein complexes we got at the end of the analyses were used as targets and a drug search was conducted using DrugBank. Results show protein complexes corresponding to immune cell functioning, inflammation, disrupted endoplasmic reticulum associated protein degradation (ERAD) pathway, etc were consistent with viral infections. The associated drugs identified from the database agrees to the ongoing research that further supported our findings. Drugs like

Acetaminophen, Sodium fluoride, Chlorhexidine, Ticagrelor and Vidarabine have been identified and suggested by a few research studies while not much data on the efficacy of Selenoproteins have been reported.

We identify two potential elements that might have reduced the accuracy of the protein selection process from network propagation. Firstly, in all three methods we used, one or more thresholds were set to select the proteins of interest. However, these thresholds were chosen empirically, not validated by any other databases or supported by other research results. Although our selected proteins are found to be correlated with SARS-CoV-2 target, a testing set that helps choose the best threshold will further optimize our protein identification process. In addition, for the guilt-by-association algorithm, protein selection was determined based on the number of Krogan protein linkages only. The ratio of Krogan proteins in the neighborhood of a specific lung protein (number of Krogan proteins over all proteins connecting to it) can be regarded as another important factor to consider.

For future work, we can explore different methods for network propagation including regularized laplacian [29]. Then a training and corresponding test set from PPI between lung tissue and a well-studied virus can be built to actively compare the performance of different propagation techniques. In addition, deep learning methods, including artificial neural networks, can also be applied to identify proteins affected by SARS-CoV-2 by looking at human-virus domain-domain association, protein functional groups and amino acid sequences [30,31]. Protein complexes are available in multiple datasets and a more comprehensive set of complexes (for example CORUM [36]) would lead to better enrichment.

In addition to the lung tissue studied in this paper, more tissue types can be analyzed in order to have a better understanding of the disease. The analysis done to create networks of the lung tissue complexes and the Krogan proteins could be applied to a different body tissue to understand the hierarchical interactions between tissues affected by SARS-CoV-2. For example, the whole brain could be analyzed to determine its affected pathways and compare the results with the lung tissue. Current research reports neurological symptoms of coronavirus include altered states of consciousness, abnormal wakefulness after the cessation of sedation, confusion, agitation, and white matter and intracerebral lesions [37]. It may be beneficial to observe how the virus interacts with proteins in the brain, and the subsequent tissue interactions for possible treatment methods. Likewise, the tissue locations shown to be affected by the Krogan proteins from the functional enrichment analyses can also undergo analysis, including reproductive organs [36] and epithelium [42]. The analysis of more tissue types could get a better overview of the protein complexes and a more accurate and wider range of drugs could have been presented. It might also be interesting to study the efficacy of selenium and selenoproteins as therapeutic agents for COVID. They are strong antioxidants, anti-inflammatory agents that promote immune cell function and have antiviral effects. They also tend to increase immunogenicity of vaccines which makes it a good therapeutic candidate.

TEAM CONTRIBUTIONS

Divya Venkat - Protein complex data download and filtering. Performing fisher test of complexes against the subnetworks. Visualization of network propagation and random walk with restart.

Aishwarya Deengar - Random Walk with Restart Algorithm, Protein function enrichment for RWR. Conducted protein complex-drug analysis

Nicole Carr - Network diffusion, guilt-by-association results, and protein functional enrichment analysis

Tian Wang - Guilt by Association and Network diffusion; visualization of guilt-by-association

REFERENCES

1. Wiersinga WJ, Rhodes A, Cheng AC, Peacock SJ, Prescott HC. Pathophysiology, Transmission, Diagnosis, and Treatment of Coronavirus Disease 2019 (COVID-19): A Review. *JAMA*. 2020;324(8):782–793. http://doi:10.1001/jama.2020.12839
2. Gordon, D.E., Jang, G.M., Bouhaddou, M. *et al.* A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* 583, 459–468 (2020). https://doi.org/10.1038/s41586-020-2286-9
3. Zhou, Y., Hou, Y., Shen, J. *et al.* (2020). Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. *Cell Discov* 6(14). https://doi.org/10.1038/s41421-020-0153-3
4. Gordon, D.E., Jang, G.M., Bouhaddou, M. *et al.* (2020). A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* 583, 459–468. https://doi.org/10.1038/s41586-020-2286-9
5. Messina, F., Giombini, E., Agrati, C. *et al.* COVID-19: viral–host interactome analyzed by network based-approach model to study pathogenesis of SARS-CoV-2 infection. *J Transl Med* 18, 233 (2020). https://doi.org/10.1186/s12967-020-02405-w
6. Wendt L, Brandt J, Bodmer BS, et al. The Ebola Virus Nucleoprotein Recruits the Nuclear RNA Export Factor NXF1 into Inclusion Bodies to Facilitate Viral Protein Expression. *Cells*. 2020;9(1):187. Published 2020 Jan 11. doi:10.3390/cells9010187
7. Uddin MH, Zonder JA, Azmi AS. Exportin 1 inhibition as antiviral therapy [published online ahead of print, 2020 Jun 20]. *Drug Discov Today*. 2020;25(10):1775-1781. doi:10.1016/j.drudis.2020.06.014
8. Noncanonical Role of FBXO6 in Regulating Antiviral Immunity. Xiaohong Du, Fang Meng, Di Peng, Zining Wang, Wei Ouyang, Yu Han, Yayun Gu, Lingbo Fan, Fei Wu,

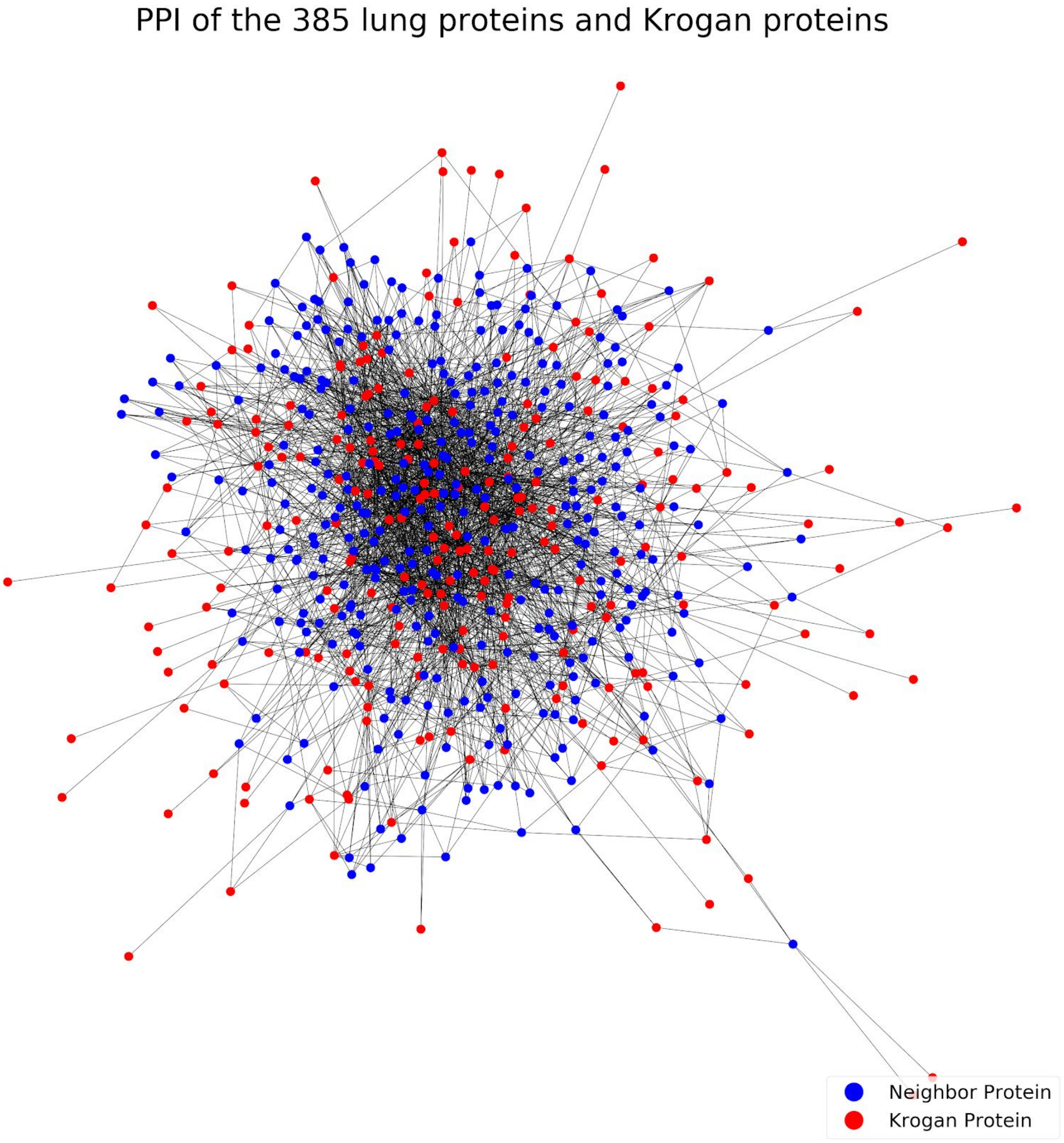
Xiaodong Jiang, Feng Xu, F. Xiao-Feng Qin. The Journal of Immunology August 15, 2019, 203 (4) 1012-1020; DOI: 10.4049/jimmunol.1801557

9. ELAVL1 Primarily Couples mRNA Stability with the 3'UTRs of Interferon Stimulated Genes. Katherine Rothamel, Sarah Arcos, Byungil Kim, Clara Reasoner, Neelanjan Mukherjee, Manuel Ascano. bioRxiv 2020.08.24.263418; doi: https://doi.org/10.1101/2020.08.24.263418
10. Szklarczyk D, Gable AL, Lyon D, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res*. 2019;47(D1):D607-D613. doi:10.1093/nar/gky1131
11. Yan K, Gao LN, Cui YL, Zhang Y, Zhou X. The cyclic AMP signaling pathway: Exploring targets for successful drug discovery (Review). *Mol Med Rep*. 2016;13(5):3715-3723. doi:10.3892/mmr.2016.5005
12. National Center for Biotechnology Information (2020). PubChem Compound Summary for CID 1983, Acetaminophen. Retrieved December 14, 2020 from https://pubchem.ncbi.nlm.nih.gov/compound/Acetaminophen.
13. Hutagalung AH, Novick PJ. Role of Rab GTPases in membrane traffic and cell physiology. *Physiol Rev*. 2011;91(1):119-149. doi:10.1152/physrev.00059.2009
14. Omarjee L, Meilhac O, Perrot F, Janin A, Mahe G. Can Ticagrelor be used to prevent sepsis-induced coagulopathy in COVID-19?. *Clin Immunol*. 2020;216:108468. doi:10.1016/j.clim.2020.108468
15. Jun Hoseki, Ryo Ushioda, Kazuhiro Nagata, Mechanism and components of endoplasmic reticulum-associated degradation, *The Journal of Biochemistry*, Volume 147, Issue 1, January 2010, Pages 19–25, https://doi.org/10.1093/jb/mvp194
16. Zhang J, Saad R, Taylor EW, Rayman MP. Selenium and selenoproteins in viral infection with potential relevance to COVID-19 [published online ahead of print, 2020 Sep 10]. *Redox Biol*. 2020;37:101715. doi:10.1016/j.redox.2020.101715
17. John C. Zinder, Christopher D. Lima. Targeting RNA for processing or destruction by the eukaryotic RNA exosome and its cofactors, *Genes & Dev.* 2017. 31: 88-100. doi:10.1101/gad.294769.116
18. National Center for Biotechnology Information. PubChem Compound Summary for CID 5152, Salmeterol. https://pubchem.ncbi.nlm.nih.gov/compound/Salmeterol. Accessed Dec. 14, 2020.
19. Strambio-De-Castillia, C., Niepel, M. & Rout, M. The nuclear pore complex: bridging nuclear transport and gene regulation. *Nat Rev Mol Cell Biol* 11, 490–501 (2010). https://doi.org/10.1038/nrm2928
20. Gil C, Ginex T, Maestro I, et al. COVID-19: Drug Targets and Potential Treatments. *J Med Chem*. 2020;63(21):12359-12386. doi:10.1021/acs.jmedchem.0c00606
21. Carlton J. The ESCRT machinery: a cellular apparatus for sorting and scission. *Biochem Soc Trans*. 2010;38(6):1397-1412. doi:10.1042/BST0381397

22. National Center for Biotechnology Information. PubChem Compound Summary for CID 21704, Vidarabine. https://pubchem.ncbi.nlm.nih.gov/compound/Vidarabine. Accessed Dec. 14, 2020.
23. Cadegiani, F.A. Repurposing existing drugs for COVID-19: an endocrinology perspective. *BMC Endocr Disord* 20, 149 (2020). https://doi.org/10.1186/s12902-020-00626-0
24. Horby P, Lim WS, et al (RECOVERY Collaborative Group). Dexamethasone in Hospitalized Patients with Covid-19 - Preliminary Report [published online ahead of print, 2020 Jul 17]. N Engl J Med. 2020;https://doi.org/10.1056/NEJMoa2021436.
25. Murali TM, Wu C-J, Kasif S. The art of gene function prediction. Nature Biotechnology. 2006;24(12):1474-1475. doi:10.1038/nbt1206-1474
26. Cowen L, Ideker T, Raphael BJ, Sharan R. Network propagation: a universal amplifier of genetic associations. Nature Reviews Genetics. 2017;18(9):551-562. doi:10.1038/nrg.2017.38
27. https://learn.bu.edu/ultra/courses/_71219_1/cl/outline
28. Yates, A., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M., Armean, I., Azov, A., Bennett, R., Bhai, J., Billis, K., Boddu, S., Marugán, J., Cummins, C., Davidson, C., Dodiya, K., Fatima, R., Gall, A., Giron, C., Gil, L., Grego, T., Haggerty, L., Haskell, E., Hourlier, T., Izuogu, O., Janacek, S., Juettemann, T., Kay, M., Lavidas, I., Le, T., Lemos, D., Martinez, J., Maurel, T., McDowall, M., McMahon, A., Mohanan, S., Moore, B., Nuhn, M., Oheh, D., Parker, A., Parton, A., Patricio, M., Sakthivel, M., Abdul Salam, A., Schmitt, B., Schuilenburg, H., Sheppard, D., Sycheva, M., Szuba, M., Taylor, K., Thormann, A., Threadgold, G., Vullo, A., Walts, B., Winterbottom, A., Zadissa, A., Chakiachvili, M., Flint, B., Frankish, A., Hunt, S., IIsley, G., Kostadima, M., Langridge, N., Loveland, J., Martin, F., Morales, J., Mudge, J., Muffato, M., Perry, E., Ruffier, M., Trevanion, S., Cunningham, F., Howe, K., Zerbino, D. and Flicek, P., 2020. Ensembl 2020.
29. Law JN, Akers K, Tasnina N, et al. Identifying Human Interactors of SARS-CoV-2 Proteins and Drug Targets for COVID-19 using Network-Based Label Propagation. arXiv.org. https://arxiv.org/abs/2006.01968. Published June 22, 2020.
30. Barman RK, Saha S, Das S. Prediction of Interactions between Viral and Host Proteins Using Supervised Machine Learning Methods. PLoS ONE. 2014;9(11). doi:10.1371/journal.pone.0112034
31. Shen J;Zhang J;Luo X;Zhu W;Yu K;Chen K;Li Y;Jiang H; Predicting protein-protein interactions based only on sequences information. Proceedings of the National Academy of Sciences of the United States of America. https://pubmed.ncbi.nlm.nih.gov/17360525/
32. Basha O, Barshir R, Sharon M, et al. The TissueNet v.2 database: A quantitative view of protein-protein interactions across human tissues. Nucleic Acids Research. 2016;45(D1). doi:10.1093/nar/gkw1088

33. Spirin V, Mirny LA. Protein complexes and functional modules in molecular networks. Proceedings of the National Academy of Sciences. 2003;100(21):12123-12128. doi:10.1073/pnas.2032324100
34. Akhter Y, Hussain R. Protein-protein complexes as targets for drug discovery against infectious diseases. Advances in Protein Chemistry and Structural Biology. 2020:237-251. doi:10.1016/bs.apcsb.2019.11.012
35. Giurgiu M, Reinhard J, Brauner B, et al. CORUM: the comprehensive resource of mammalian protein complexes-2019. Nucleic Acids Res. 2019;47(D1):D559-D563. doi:10.1093/nar/gky973
36. Li R, Yin T, Fang F, et al. Potential risks of SARS-CoV-2 infection on reproductive health. *Reprod Biomed Online*. 2020;41(1):89-95. doi:10.1016/j.rbmo.2020.04.018
37. Kremer, S., Lersy, F., de Sèze, J., Ferré, J. C., Maamar, A., Carsin-Nicol, B., ... & Rafiq, M. Brain MRI findings in severe COVID-19: a retrospective observational study. *Radiology*, 2020;297(2): E242-E251.
38. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc. 2009;4(1):44-57. [PubMed]
39. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37(1):1-13.
40. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* 2003 Nov; 13(11):2498-504.
41. Merico D, Isserlin R, Stueker O, Emili A, Bader GD, Enrichment Map: A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation, *PLoS One*. 2010 Nov 15;5(11):e13984.
42. Brann, D., Tsukahara, T., Weinreb, C., Logan, D. W., & Datta, S. R. Non-neural expression of SARS-CoV-2 entry genes in the olfactory epithelium suggests mechanisms underlying anosmia in COVID-19 patients. *BioRxiv*. 2020.

APPENDIX



PPI of the 385 lung proteins and Krogan proteins

Figure S1: Network visualization of 385 lung proteins that are implicated in guilt-by-association algorithm

Additional Supplementary materials are attached in a zipped folder.
The folder contents include

1. Jupyter notebooks for the three methods and visualizations.
2. Excel file showing the enriched protein complexes obtained from each method
3. Text files of DAVID functional enrichment annotation clustering results for each method