UNIVERSITÀ DI TRENTO

Data mining approaches to COVID-19 and influenza virus for drug repurposing

EMMA BUSARELLO, MATTEO POZZI, FRANCESCO PENASA and GABRIELE BERRERA

Professors: Enrico Blanzieri, Stefania Pilati, Valter Cavecchia, Toma Tebaldi

October 25, 2020

Abstract:

SARS-CoV-2 is the coronavirus responsible for the COVID-19 pandemic which already caused 46.2 million cases and more than 1 million death worldwide. At the present there are no effective treatment for this disease.

For this reason, the aims of this research are the investigation of protein interactions of SARS-CoV-2 and the comparison between COVID-19 and H1N1 influenza virus for possible drug repurposing. The development of the project is based on three different parts: network based analysis, enrichment analysis and drug repurposing. From Gordon et al. study [1], 332 virus-interacting proteins, related to COVID-19, have been downloaded and then expanded as OneGenE on the Fantom 5 dataset [2][3], while gene set related to influenza virus (H1N1) was downloaded from EnrichR library [4][5]. Genes were extracted according to the rank value, and the most relevant genes were then expanded using NES^{2}RA algorithm [6].

By applying gene network-based analysis, for COVID-19 genes and H1N1 genes, through Networkx Python pgk, Gephi [7] for visualization, and STRING API [8] for validation, we investigate the relationships between genes. We then compared the two networks in order to extract common genes. Subsequent functional enrichment analysis was performed using COVID-19 genes, to explore genes and related pathways that can be relevant for this viruses, and then also with common genes, allowing us to investigate significant pathways and biological processes involved in the spread of the two infections.

Thus, our purpose is to investigate the relation of these genes with possible drugs, already available for influenza virus, for drug repurposing. To perform this kind of analysis, we used OpenTarget Platform [9] and DGIdb [10], and for a further validation we compare the results obtained with a set of targets available also for COVID-19. This lead us to the identification of new potential targets and to validation of known aspect of the current pandemic as well as to a novel analysis and comparison between two viruses that affects human health.

Introduction

The novel SARS-CoV-2 is the strain of coronavirus that is responsible for the COVID-19 pandemic, causing severe respiratory disease. It is a single-stranded RNA virus and it has four structural proteins that constitute the spike, envelope, membrane and nucleocapsid. The spike protein is the one responsible for the attachment and fusion with the membrane of the human cells and ACE2 is its cellular target.

COVID-19 has affected currently a population of 46,2 million people and caused more than 1 million death. At the moment there are not any antiviral drugs to this pathology, vaccines are close to be approved around the world but unfortunately, the scientific community has little knowledge of the molecular details of SARS-CoV-2 infection. Additionally, due to the incoming winter season, also other factor will become relevant, such as common symptoms related to the well known flue. One important aspect is that these symptoms can cause respiratory disease and can be transmitted by contacts and droplets.

H1N1 influenza virus is a sub-type of influenza A and is the cause of the 2009 Swine flue pandemic as well as the 1918 flue pandemic. It belongs to the family of orthomyxovirus which contains the glycoproteins: heamagglutinin (H) and neuramidase (N). Influenza A can be divided into 16 types (H1-H16) and nine subtypes (N1-N9). The heamagglutinin proteins are the responsible for the receptor-mediated endocytotsis of the virus inside the cytoplasm of the host cells by recognition of the sialic acid receptors of host cells. The genome of Influenza A virus contains eight pieces of singles-strand RNA, which can be translated into ten viral proteins, such as polymerase, nuclear export protein [11]. Similarly to COVID-19, apparently, also Swine influenza originated from a zoonosis, since it is a typical respiratory infection that occurs in pigs [12].

Even if the molecular mechanism of the infection and the cellular target are different in the two viruses, a comparison between the two protein-protein interactions networks could give insight into new possible ways of treating a newly and partially understood disease by exploiting an older one.

Data preparation

Data relative to COVID-19 were downloaded from a previous study: using affinity-purification mass spectrometry were identified 332 high-confidence protein-protein interactions between SARS-CoV-2 proteins and human proteins[1]. These virus-interacting proteins have been expanded as OneGenE on the Fantom 5 dataset. The Fantom dataset has been created with the CAGE technology, which identifies all the transcriptional starting sites (TSS) for a gene. For this reason, 332 genes correspond to 1627 gene isoforms.The list of expanded genes were filtered according to the relative frequency, selecting 95% as threshold.

Data relative to influenza virus were collected from EnrichR in "libraries" section, where "Virus_Perturbation_from_GEO_down" and "Virus_Perturbation_from_GEO_up" gene set were downloaded. In order to make a comparable analysis with human COVID-19 genes, in the two downloaded files we selected one GEO dataset containing only human related genes associated with H1N1 virus (GSE40844), in which the authors compared the differential expression of genes in human lung cell lines affected with influenza and a control. We then combined the two files obtaining a unique list of genes, ranked according to a specific value, which is reported in EnrichR as the combination of p-value, q-value and z-score.

In order to select an appropriate threshold for filtering these genes list, we plotted according to their significance. A graphical visualization of the ranked list allow us to select the most 100 relevant genes.

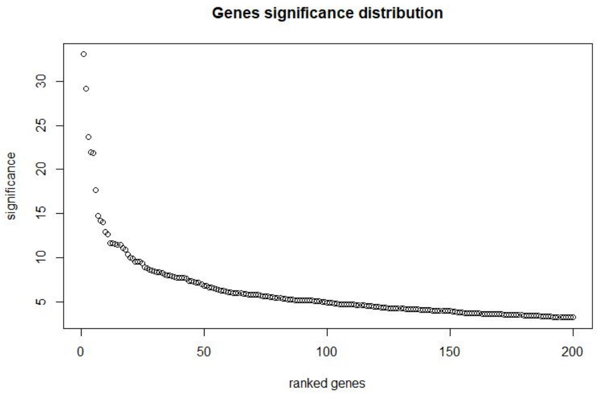


Figure 1: Genes significance distribution

Validation of the data

The 332 COVID-19 genes were first compared to gene sets available on EnrichR library section ("COVID-19_Related_Gene_Sets"). This library collects gene sets related to COVID-19 thanks to the contributions of several researches, including the gene set produced by Gordon et al. study [1]. For this reason we excluded these genes for the validation procedure. By matching the two list of genes we were able to detect 51 genes in common, implying that they are already found in other COVID-19 studies, while most of the 332 genes have not been detected in other studies yet.

Workflow

To reach the aims of this project, on one side the investigation of protein-protein interactions of COVID-19 in comparison with H1N1 virus, and on the other, a drug repurposing process, we performed a set of different analysis:

- NES^{2}RA: espansion of 100 significant genes of H1N1
- Network based analysis: Networkx Pyhotn pgk, Gephi and PC-algoritm
- Enrichment analysis: EnrichR, g:Profiler [13] and STRING
- OpenTarget Plaform (version 3.20.0) and DGidb drug repurposing

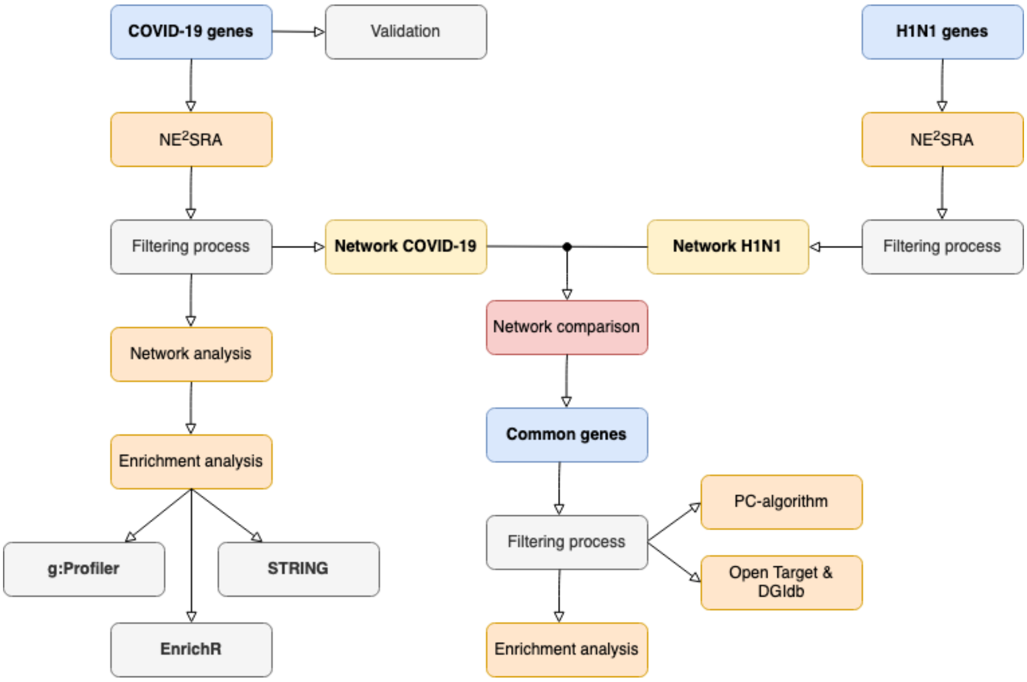


Figure 2: Analysis workflow

NES^{2}RA: H1N1 genes expansion

Gene network expansion is essential to understand all possible interactions among genes related to a particular network. NES^{2}RA is a method that use PC-algorithm in order to find causal relationships among genes. We run this method on the BOINC platform in order to expand our list of 100 H1N1 related genes, thanks to the computational power provided by gene@home volunteers. NES^{2}RA was run with the Full Fantom database and the default parameters on the BOINC system.

Since we have extracted 100 important genes, given the p-value present in the ErichR dataset, our objective is to discover if some of the genes in the H1N1 expansion also belongs to SARS-COV2 gene expansion. We expanded 278 genes isoforms de-novo while 260 were already present in the gene@home PC-IM history. The results of the expansion process retrieved two files, one with the genes of the expanded graph and the second with their interactions. For subsequent analysis we considered only the .expansion file, that has genes ranked according to their relative frequency. As we did for SARS-COV2 gene set expansion, firstly we filtered out all the genes with a frequency below 95% frequency and then we built a protein-protein interaction network.

By performing the series of analysis we have mentioned, we expect to point out relevant genes and significant pathways involved in the spreading and development of COVID-19, and additionally, finding possible correlations with H1N1 virus.

Network-Based Analysis

COVID-19 and H1N1 networks construction

To represent the interactions resulted from the NES^{2}RA expansion, we decided to build a network for the two different viruses. For doing this, we took advantage of the Python package Networkx, and starting from the seed genes as nodes, we began to add the linkages to those genes that were present into the expansion lists. Beyond the filter previously applied on the relative frequency (greater than 95%), we decided to subject the data to another type of selection. To simply explain it, let's suppose we have two different genes, *gene 1* and *gene 2*: to have a connection between those two genes, at least one half of the isoforms relative to *gene 1* have to include *gene 2* in their expansion list. The reasoning behind this choice is driven by the fact that, since we considered the most connected part of the network to establish the most significant genes, we notice that genes that encodes for a larger number of isoforms result advantaged.

Reduction of the networks

Once the two networks were built, they were visualized and analysed using Gephi, in order to determine their core components. The network relative to COVID-19 was composed of more than 8000 genes and they seemed to be very connected, so we decided to apply a strong cut, filtering out all the nodes with a degree less than 10. The resulting network was composed of 213 connected genes, which were then subjected to further analysis.
As regards H1N1, since the expansion involved a smaller number of genes (100 genes, despite the 332 for the COVID-19), the resulting network was composed of about 3000 genes. In this case, to find the core of the graph, was sufficient to select largest connected component and remove all the leaves, resulting in a subnetwork composed of 600 genes.

Networks validation

Once the two networks relative to H1N1 and COVID-19 were reduced with Gephi and the core components were found, we tried to validate them using the data stored in the STRING-DB. We created a Python script that, using the STRING API for accessing the database, compares the data with our networks and tries to estimate a level of similarity. We basically computed the proportion of pairwise linked genes that were present both in our graph and in STRING-DB. Both for COVID-19 and H1N1 network, we found only a limited part of the network to be in common with STRING: around the 27% for COVID-19 and 20% for H1N1. This result doesn't necessarily mean that the graphs we built are incorrect, indeed not all the known interactions are present on the STRING database.

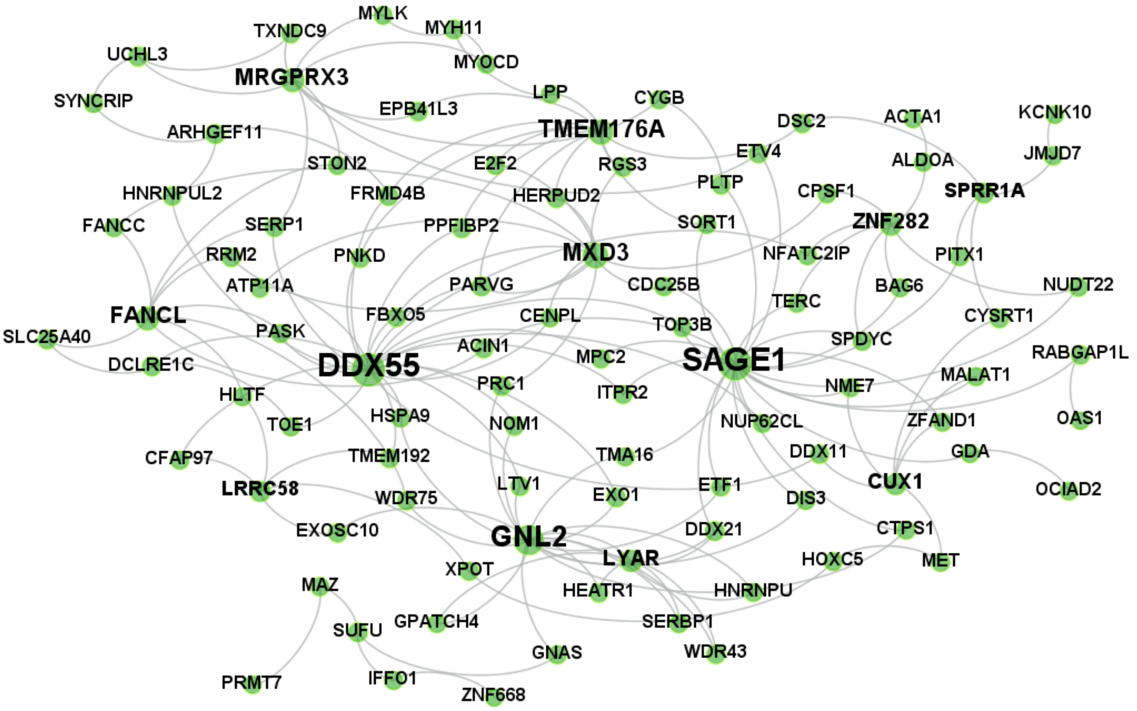


Figure 3: Filtered network composed by 100 genes that are present in both COVID-19 and H1N1 networks

Comparison of COVID-19 and H1N1 networks

In order to find out what are the possible similarities between the two viruses, we made a comparison of the two networks, searching for the genes that are in common, and computing the intersection of the two graphs.

Comparing the nodes of the two networks we discovered 2045 shared genes. In order to reduce this list of genes to a significant subset, we decided to project them onto the H1N1 graph, for then reducing the network and determine which is the core component. For doing this, we used Gephi to selected the largest connected component of the network, and then we filtered out all the nodes with a degree less than 2, in order to remove the leaves of the graph. In Figure 3 we reported the resulting network, composed of 100 nodes. This list of those 100 genes was subjected to a functional enrichment analysis and a search for possible drug targets.

To perform a more accurate comparison of the two network, instead of selecting only the genes that are in common, we decided to compute the intersection of the two graphs. Using the Python package Networkx, we basically selected all the couples of genes that we found to be directly connected in both the two networks, resulting in a subnetwork composed by only 17 nodes.

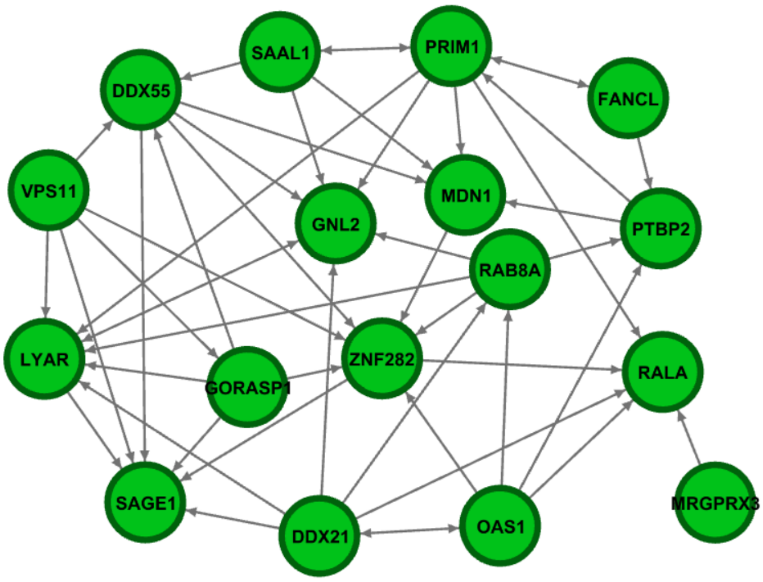


Figure 4: Subnetwork created by running the Pc-Algorithm on the 17 genes resulted from the intersection of the two graphs.

To better understand the possible relationship between those genes and the directionality of their dependencies, we ran the PC-Algorithm using the Gaussian Conditional Independence test, with an alpha level of 0.01. The resulting network is shown in Figure 4, represented this time as a directed graph. This 17 genes were then further analysed and compared with the literature, in order to find some biologically meaningful information.

Functional Enrichment Analysis

The next step in our project was to perform some enrichment analysis on the lists of relevant genes, that were extrapolated from the analysis of the different networks. The aim was to find biological information present on the literature, that can be linked to our sets of genes.

Analysis of COVID-19 most significant genes

A first functional enrichment analysis was performed on the 213 most relevant COVID-19 genes using g:Profiler (version e101_eg48_p14_baf17f0), with Benjamini-Hochberg FDR multiple testing correction method applying significance threshold of 0.05. We also performed the functional enrichment analysis using STRING and EnrichR online version for validation. STRING and g:Profiler enrichment gave almost the same results, while EnrichR provided a smaller number of enriched genes, but still significant with a p-value less than 0.05.

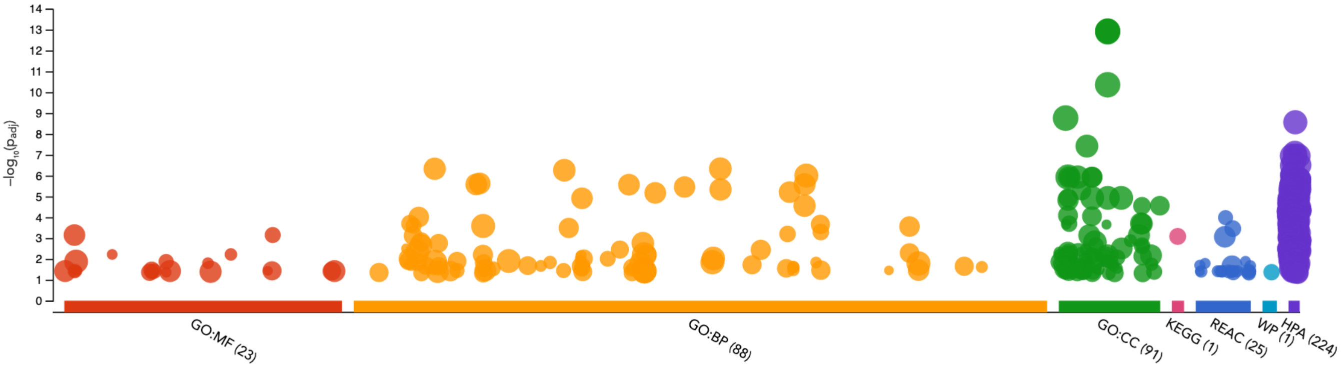
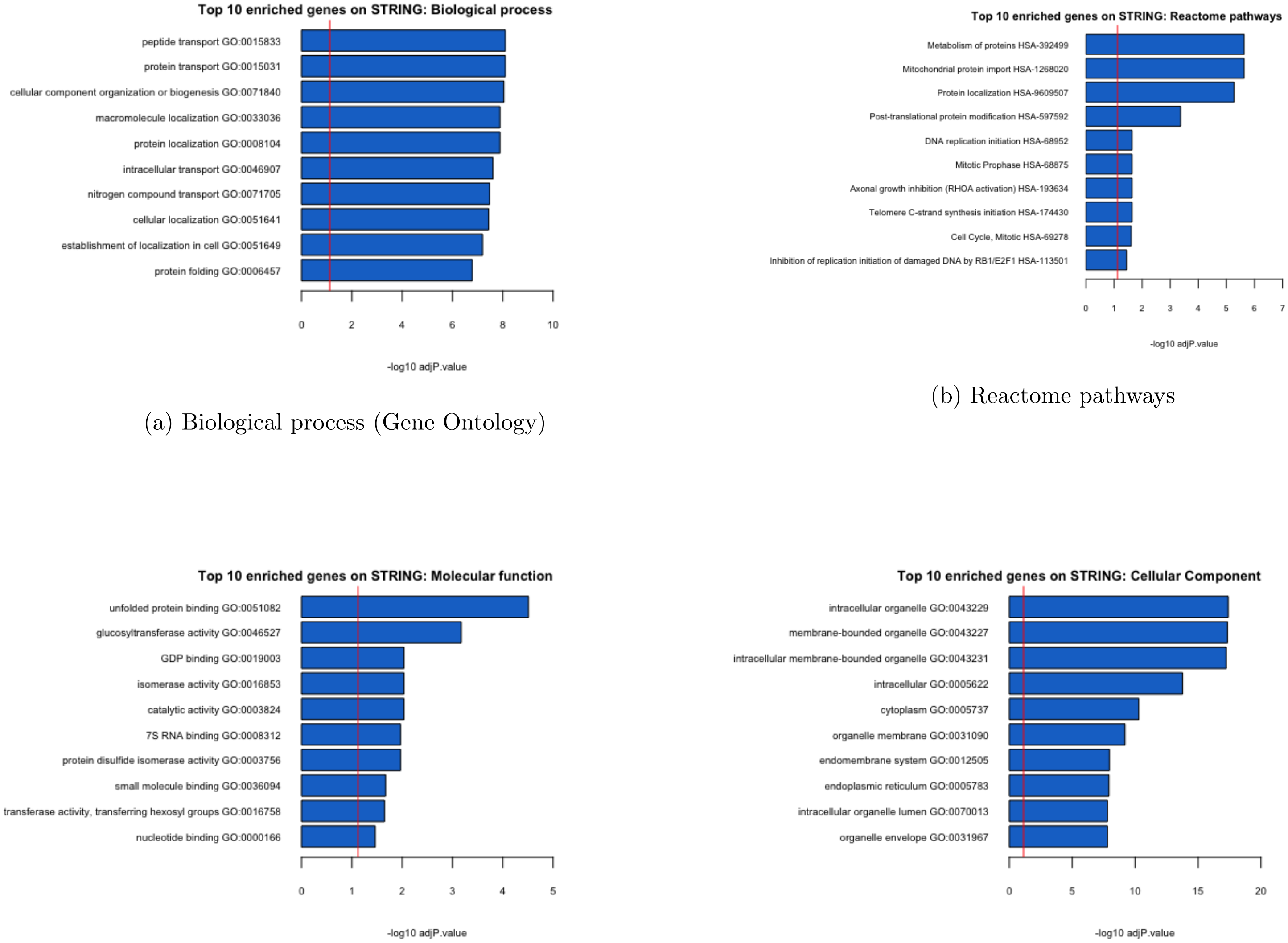


Figure 5: g:Profiler significant terms with a threshold of 0.05 FDR

We decided to analyze, through a literature research, the results produced by STRING and EnrichR, since g:Profiler retrieved similar enriched terms produced by STRING, implying a validation of the results.



(a) Biological process (Gene Ontology)

(b) Reactome pathways

(c) Molecular function (Gene Ontology)

(d) Cellular Component (Gene Ontology)

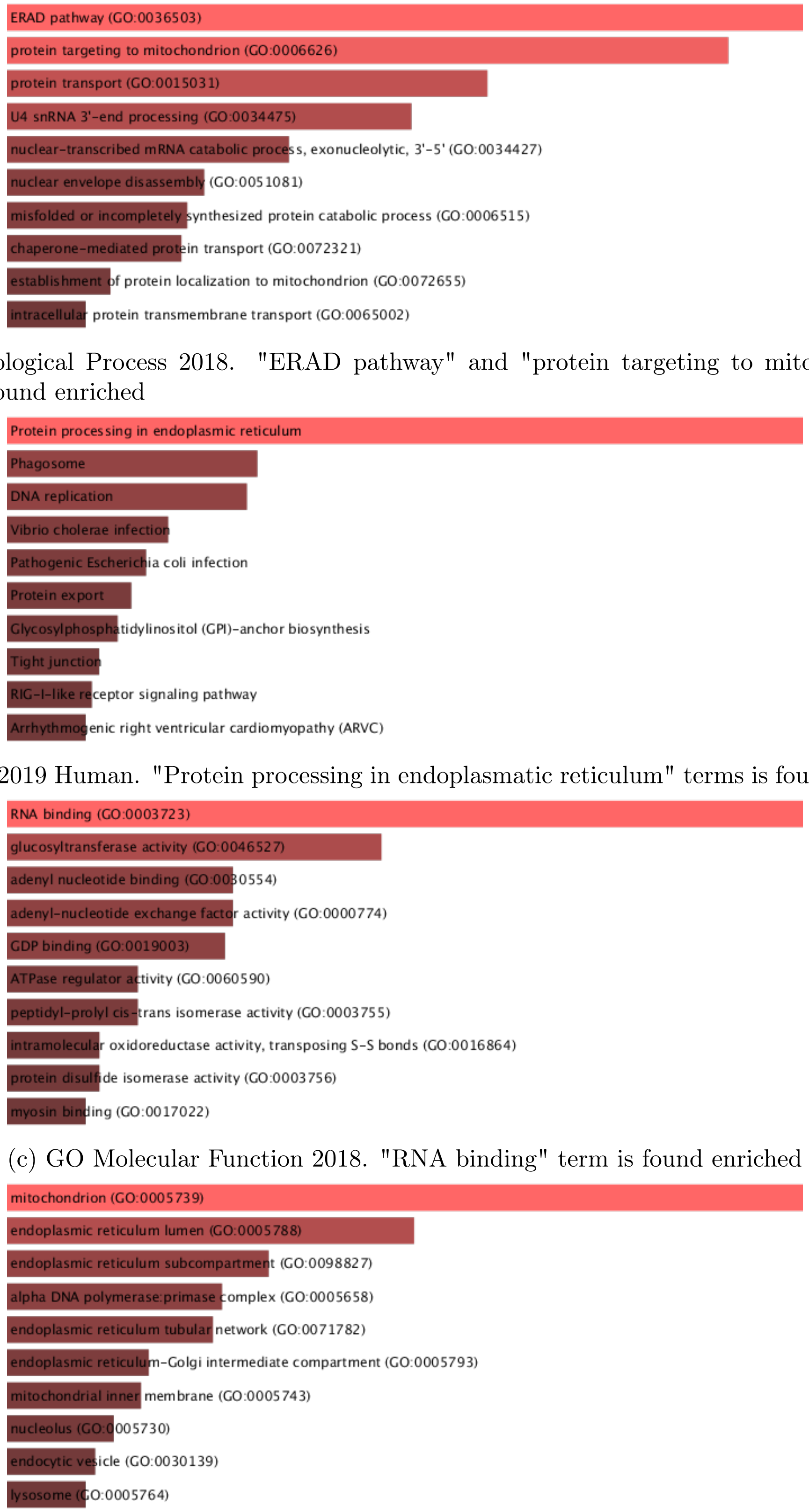
Figure 6: Barplot of the 10 most enriched terms on STRING online tool

In order to identify possible pathways enriched in our list of interacting coronavirus proteins, we retrieved also associated KEGG annotations using STRING online version, ranking terms according to Benjamini-Hochberg FDR and applying significance threshold of 0.05.

This analysis identified only one significantly enriched KEGG pathway: "protein processing in endoplasmic reticulum" (hsa04141). In addition, gene ontology results showed that, in reactome pathways and biological process, enriched genes are involved in metabolism, transport and post-translation modification of proteins. Terms related to cell cycle were also found enriched. In molecular function and cellular component annotations, it is possible to observe the "unfolded protein binding" term and several terms related to intracellular compartment. Cellular component enrichment retrieved also organelle terms that are involved in intracellular trafficking between the nucleus and cytoplasm, that is an essential cellular process, and these interactions most likely contribute to the entry, assembly and/or secretion of viral particles. The high presence of genes in binding process, catalytic activity and cell cycle process, is due to the fact that are necessary for basic life functionality.

To have an additional comparison, we also performed enrichment analysis using EnrichR online tool, showning similar results gave by STRING, but it provided less enriched genes ranked according to p-value. Significant terms, with a p-value less than 0.05, are showed with a light red color. In GO biological process 2018 of EnrichR "endoplasmic-reticulum-associated protein degradation (ERAD) pathway" is found enriched. This terms, together with the "protein processing in endoplasmatic reticulum", in the KEEG annotation, and also with "unfolded protein binding" from STRING, are in lines with previous studies, which show that COVID-19 replication in infected host cells may perturb protein folding in the endoplasmatic reticulum (ER), inducing a strong ER stress response. The unfolded protein response, that is the adaptive cellular response to misfolded protein, can controls this perturbation process, but continued viral proliferation may induced inflammation and cell death [14].

In GO molecular function 2018 "RNA-binding" is identified as the most enriched term. The RNA binding proteins (RBP) regulated the post-transcriptional gene regulatory network in humans and the dysregulation of RBPs have been shown to contribute significantly to altered this network in several diseases such as cancer, genetic diseases and viral infection. SARS-CoV-2 proteins can bind several human RBPs and thus, by regulating the transcribed viral RNA, RNA protein binding could contribute to virus assembly and export and could therefore be implicated as therapeutic targets [15].



(a) GO Biological Process 2018. "ERAD pathway" and "protein targeting to mitochondrion" terms are found enriched

(b) KEGG 2019 Human. "Protein processing in endoplasmatic reticulum" terms is found enriched

(c) GO Molecular Function 2018. "RNA binding" term is found enriched

(d) GO Cellular Component 2018. "Mitochondrion" term is found enriched

Figure 7: Bar graph visualization of the 10 most enriched terms on EnrichR online tool. Significant enriched terms, with a p-value less than 0.05, are showed with a light red color

In GO cellular component 2018, the most enriched term is mitochondria: it is not surprising. In fact, it has been shown that viruses actively interfere with mitochondrial pathways to impede mitochondrial anti-viral signalling mechanisms [16].

Enrichment analysis of common genes

To explore the biological meaning of the subset of 100 common genes among COVID-19 and H1N1, a functional enrichment analysis was performed, with the same tool used for COVID-19 genes. Interesting results were showed by EnrichR online tool, were enriched terms are ranked according to p-value. Significant enriched terms, with a p-value less than 0.05, are showed with a light red color (Fig.8, Fig.9, Fig.10, Fig.11).

Results In GO Biological Process 2018 the most enriched terms are related to ribosome biogenesis and regulation of transcription, processes involved in viral replication. In particular, the quantity and the quality of ribosome are fundamental in controlling viral protein synthesis. In some recent studies, have been demonstrated that ribosome biogenesis factor (RBFs) and ribosomal proteins (RPs) regulate the selective translation of viral transcripts. This direct interaction with viral particles can guide to a reasonable new potential antiviral agent, by targeting ribosome production and function [17].

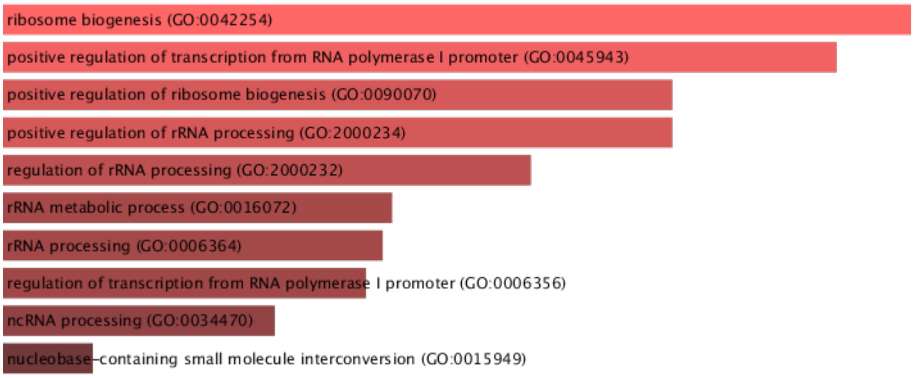


Figure 8: GO Biological Process 2018

Molecular functional enrichment retrieves only one significant term, "RNA binding", that we have found in the previous enrichment analysis of COVID-19 genes. Enrichment results in cellular component show terms related to nuclear space, in particular "nucleolus" is the most significant. All single-strand RNA viruses, as SARS-CoV-2 and H1N1 influenza virus, encode a single-strand RNA-binding nucleoprotein (NP), which role is to encapsidate the virus genome for RNA transcription, replication and packaging. According to the study of Wurm et al. [18], nucleoproteins can be found either in the cytoplasm and in a structure of the nucleus, that is the nucleolus.

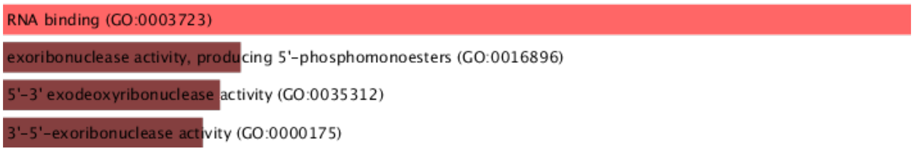


Figure 9: Go Molecular Function 2018

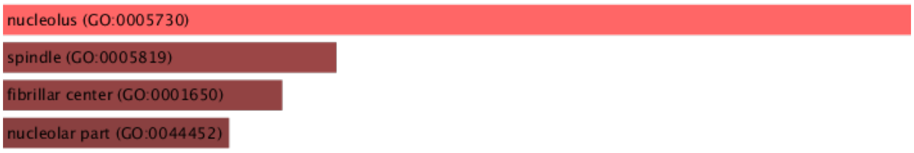


Figure 10: GO Cellular Component 2018

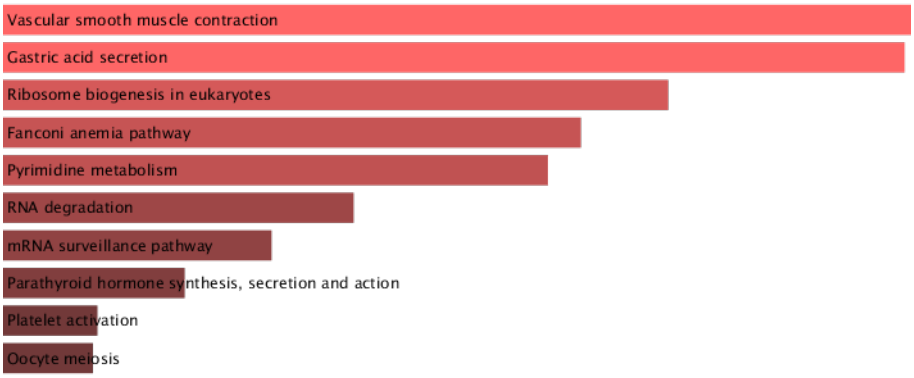


Figure 11: KEEG 2019 Human

One interesting result is given by the KEEG pathways enrichment, where the top enriched term is related to vascular smooth muscle activity. We then explored genes involved in this pathway using Laverne Novus Bioinformatics tool. Laverne is a virtual research assistant, librarian and data cruncher, that allows to search unbiased, bioinformatics data related to your gene, pathway or disease of interest (www.novusbio.com/explorer). By using this tool, given as input the enriched pathway "vascular smooth muscle contraction", were retrieved connections with 10 genes including angiotensin II that is related to influenza and COVID-19. Angiotensin II plays a central role in COVID-19 diffusion, due to the fact that SARS-CoV-2 virus enters in the target cells by binding to angiotensin-converting enzyme 2 (ACE2), resulting in activation of the renin-angiotensin system (RAS), that is responsible for regulation of cardiovascular function. High level of angiotensin II induces thrombin formation and impairs fibrinolysis, features that are found to be strongly present in patients with severe COVID-19 [19].

Moreover, ACE2 plays an important role in acute lung injury induced by influenza viruses, suggesting that ACE2 still has unexpected aspect with clinical implications [20]. For these reasons, as most of the scientific community suggested, ACE2 can be a possible target, by blocking angiotensin II production with ACE inhibitors, reducing its inflammatory results and rendering the virus less infectious as it would decrease the receptors available for the SARS-CoV-2 virus to enter cells.

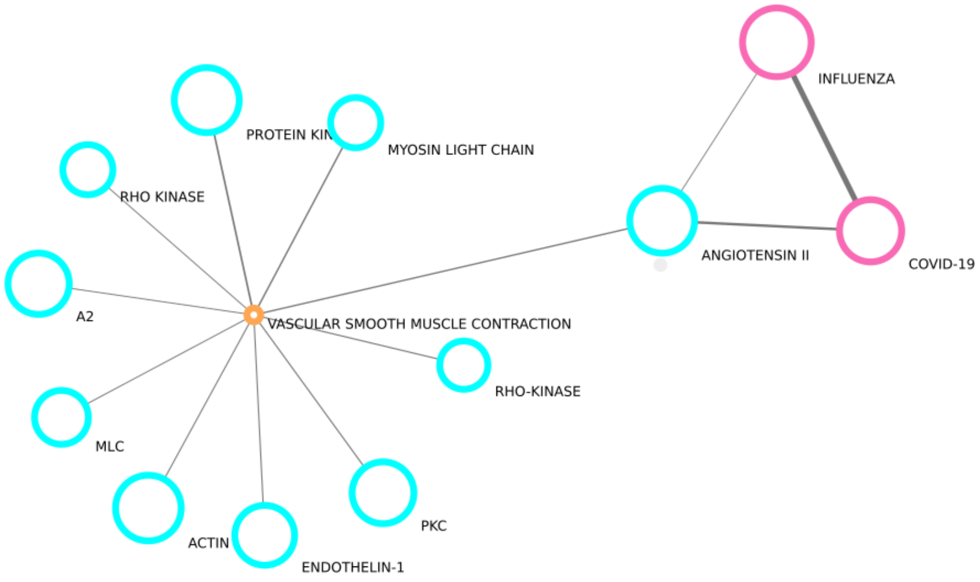


Figure 12: Vascular Smooth Muscle Contraction Pathway Laverne Bioinformatics tool visualization.

In addition to this interesting result, KEEG enrichment analysis retrieved also gastric and secretion as significant enriched pathway in the list of common genes. In fact, according to previous studies, COVID-19 virus have been found in gastrointestinal cells, in which ACE2 receptor is highly expressed. This implies that the virus can infect the gastrointestinal tract and as a consequence used by the viruses also as replication site [21]. In a a recent study of Shirey et al., was observed that gastrin-releasing peptide (GRP), that is known to mediate gastric and secretion in the gut, has been implicated in pulmonary lung inflammatory diseases, such as influenza, revealing GRP as a potentially novel target for therapeutic intervention, and maybe can be also a novel target for COVID-19 [22].

Research of Possible Targets and Drug Repurposing

In order to find possible new target and to understand which pathology are correlated with the list of 100 common genes, we performed an analysis using the Open Target online platform (version 3.20.0). Once we have extracted the common gene list we matched it with the Open Target list for H1N1. The drugs selected from the previous are compared with the Open Target relative to COVID-19 and validated if they correspond, otherwise the H1N1 drugs are proposed as now possible therapy. If instead no matches would be found, an extended drug repurposing analysis is needed, using the 213 genes related to COVID-19, used before for the enrichment analysis. For this purpose Open Target Platform would be used with all databases.

- Matches: validation through Open Target for COVID-19. Are there any correlations?
  - Yes: validation of our results
  - No: drug repurposing (possible new target for COVID-19)
- Mismatches: expand the search to all databases

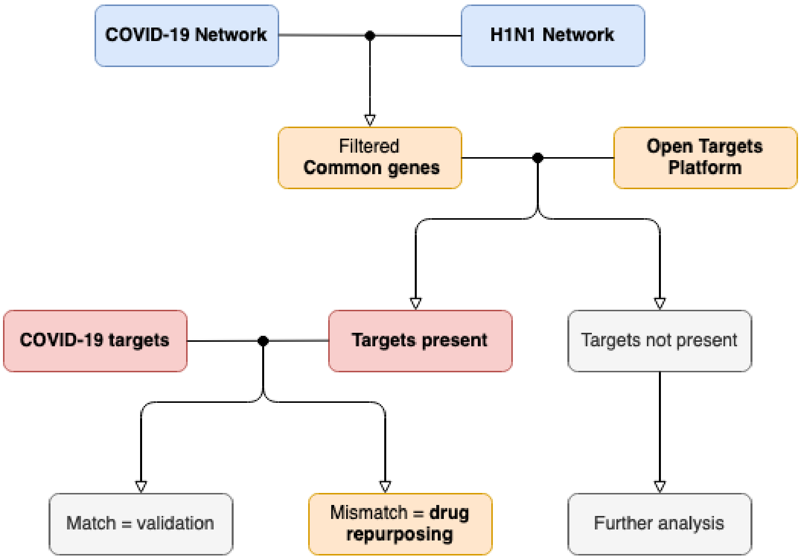


Figure 13: Process for drug repurposing

Open Target results By restricting the search to a H1N1-related target list matched with our 100 common genes, only one results is present in both lists: PLTP (phospholipid transfer protein) which mediate the transfer of phospholipids and free cholesterol from triglyceride-rich lipoproteins (low density lipoproteins or LDL and very low density lipoproteins or VLDL) into high-density lipoproteins (HDL) as well as the exchange of phospholipids between triglyceride-rich lipoproteins themselves. This target has no Drugs currently available according to Open Target neither is a target present for COVID-19 in Open Target database. We decided to search this gene also on DGIdb and we found that there is one drug, approved from FDA, that act as anticholesterolaemic agent. COVID-19 complications seems to be more severe in patients with cardiovascular disease, hypertension, and overweight/obesity and appears that lowering lipid and cholesterol levels can reduce the complacence of the infection as well as have an additional antiviral activity by indirectly reducing the concentration of ACE2 receptors and so reducing viral infectivity [23]. Since we have reached few significant results, we decided to expand the the research for COVID-19 drug repurposing to all Open Target Database, by using the 213 COVID-19 genes. Open Target Platform allows to run only 200 genes, so we selected the first 200. We also decided to use DGIdb to find more target genes among the list of 200 excluding the ones founded previously with Open Target.

SIRT5 One possible target is SIRT5. This gene encodes a member of the sirtuin family (nicotinamide adenine dinucleotide (NAD)+ dependent deacetylases) of proteins that are known to play an important role in cellular homeostasis. SIRT5 interact with Nsp14, a $3'$-$5'$ exonuclease, that is critical for coronavirus RNA synthesis by capping the viral RNAs so that they can evade immune recognition. SIRT5 inhibitors are already developed in order to trying to interfere with this process and thus with the production of infectious viral particles [24].

HMOX1 is another interesting association. It codify for the human heme oxygenase which is an enzyme that catalyzes the degradation of the heme ring. It also exhibits cytoprotective effects since excess of free heme sensitizes cells to undergo apoptosis. Recently it has been proposed the hypothesis that inducing the HO-1 and consequently increase the heme catablosism could confer antiviral activity [25].

Drug	**Target**	**Disease**	**Molecule type**
CYTARABINE	POLA1	acute myeloid leukemia	Small molecule
METFORMIN	NDUFAF2	type II diabetes mellitus	Small molecule
MYCOPHENOLATE MOFETIL	IMPDH2	chronic kidney disease	Small molecule
LENALIDOMIDE	RBX1	multiple myeloma	Small molecule
THALIDOMIDE	RBX1	Mantle cell lymphoma	Small molecule
FLUDARABINE PHOSPHATE	PRIM1	chronic lymphocytic leukemia	Small molecule
GEMCITABINE	PRIM2	extranodal nasal NK/T cell lymphoma	Small molecule
MYCOPHENOLIC ACID	IMPDH2	infection	Small molecule
THIOGUANINE	IMPDH2	adult acute lymphoblastic leukemia	Small molecule
AMINOCAPROIC ACID	PLAT	hemorrhage	Small molecule
AZACITIDINE	DNMT1	myelodysplastic syndrome	Small molecule
DECITABINE	DNMT1	myelodysplastic syndrome	Small molecule
DEXTROMETHORPHAN	SIGMAR1	sinusitis	Small molecule
COLLAGENASE CLOSTRIDIUM HISTOLYTICUM	COL6A1	Dupuytren Contracture	Enzyme
PENTAZOCINE	SIGMAR1	pain	Small molecule
CARBETAPENTANE	SIGMAR1	sinusitis	Small molecule
CLOFARABINE	PRIM1	neoplasm	Small molecule
FENFLURAMINE	SIGMAR1	obesity	Small molecule
BRODALUMAB	IL17RA	psoriasis	Antibody
POMALIDOMIDE	RBX1	immune system disease	Small molecule
OCRIPLASMIN	COL6A1	eye disease	Enzyme
VOLOCIXIMAB	ITGB1	ovarian cancer	Antibody
TROXACITABINE	PRIM1	acute myeloid leukemia	Small molecule
MOLIBRESIB	BRD4	neoplasm	Small molecule
FIRATEGRAST	ITGB1	multiple sclerosis	Small molecule
OZANEZUMAB	RTN4	amyotrophic lateral sclerosis	Antibody

Table 1: Open Target results: list of genes of COVID-19 with related drug

Gene	**Drug**	**Interaction types**
SCAP	ATORVASTATIN	
SCAP	SIMVASTATIN	
NDUFAF1	METFORMIN HYDROCHLORIDE	inhibitor
TOR1A	HALOPERIDOL	
SLC30A9	ASPIRIN	
CEP68	ASPIRIN	
SCARB1	ATORVASTATIN	
SCARB1	FENOFIBRATE	
SCARB1	RIBAVIRIN	
ADAMTS1	PRAVASTATIN	
NEK9	FOSTAMATINIB	inhibitor
SIRT5	CEFIXIME	
SIRT5	NIACINAMIDE	
SRP19	CITALOPRAM	
SRP19	SERTRALINE	
SRP19	PAROXETINE	
SRP19	FLUOXETINE	
HMOX1	SUNITINIB	
HMOX1	SORAFENIB	
HMOX1	ASPIRIN	
PPT1	LUCINACTANT	
PPT1	LAMOTRIGINE	
ACTB	ETHINYL ESTRADIOL	
ACTB	CYCLOPHOSPHAMIDE	
RHOA	PRAVASTATIN	
RHOA	SIMVASTATIN	
F2RL1	ERYTHROMYCIN	
F2RL1	DOXYCYCLINE	
F2RL1	MINOCYCLINE	
F2RL1	CLARITHROMYCIN	
F2RL1	ROXITHROMYCIN	
NSD2	MITOXANTRONE	
TBK1	FOSTAMATINIB	inhibitor
TBK1	AMLEXANOX	inhibitor

Table 2: DGIdb results: list of genes of COVID-19 with related drug

IMPDH2 It is and enzyme that plays a crucial role in the de novo synthesis of guanine nucleotides, and therefore in the regulation of cell growth. IMPDH inhibitors are currently investigated as potential antiviral drugs since it has been demonstrated that can suppress replication of a variety of merging RNA viruses. It has also been proved that IMPDH inhibitor merimepodib (MMPD) supresses SARS-CoV-2 replication in vitro [26].

Analysis of the intersection genes

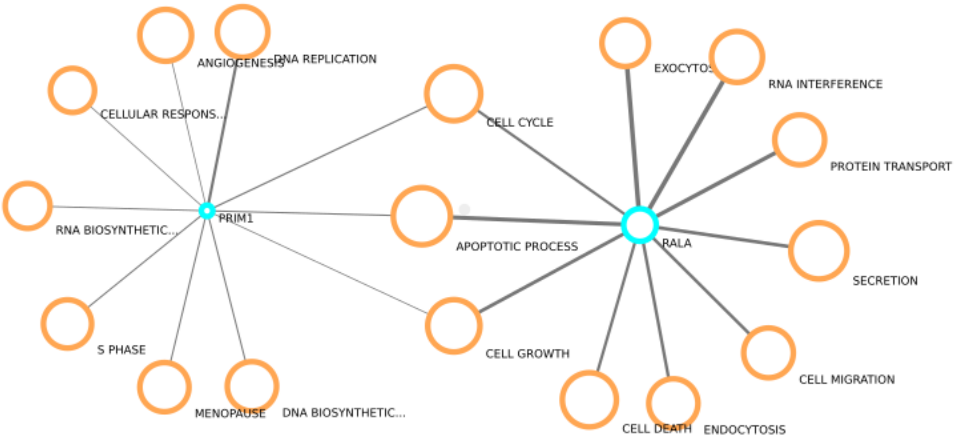
Last, we investigate the presence of potential targets for COVID-19 or H1N1 in the intersection network shown in Figure 4, by using Open Target Platform (version 3.20.0) and DGIdb (v4.2.0 - sha1 afd9f30b). From Open Target analysis one gene target was retrieved, PRIM1, that is not associated with influenza or COVID-19, but instead is related to leukemia and cancer.

Also DGIdb analysis retrieved only one target gene, that is RALA with CHEMBL384759 as associated drug. By looking at the dependencies retrieved by PC-algorithm, it is possible to observe the connections between PRIM1, RALA and the others genes in the network. PRIM1 acts on four different genes, LYAR, GNL2, MDN1 and also RALA, implying that if PRIM1 is inhibited, as a consequence there is an interference with the action of the others genes.

Drug	**Target**	**Disease**	**Molecule type**
FLUDARABINE PHOSPHATE	PRIM1	chronic lymphocytic leukemia	Small molecule
CYTARABINE	PRIM1	acute lymphoblastic leukemia	Small molecule
GEMCITABINE	PRIM1	pancreatic carcinoma	Small molecule
CLOFARABINE	PRIM1	neoplasm	Small molecule
TROXACITABINE	PRIM1	acute myeloid leukemia	Small molecule

Table 3: Open Target results on 17 genes list. PRIM1 target with associated drugs for several diseases

In addition, we explored the pathways linked to RALA and PRIM1 using Laverne Novus Bioinformatics tool. This tool allowed us to extend pathways that are associated to RALA and PRIM1 and find also possible common features.



Powered by Biovista Vizit

Figure 14: RALA and PRIM1 pathways retrieved by Laverne Novuc Bioinformatis tool

RALA and PRIM1 genes are found to be involved in the cell cycle, cell growth and apoptosis. The DNA primase polypeptide 1 (PRIM1) is responsible for synthesizing small RNA primers which play an essential role in the initiation of DNA synthesis. According to Lee et al. study [27] that inhibition of PRIM1 caused significant cell growth cycle arrest at the G2/M phase.

Discussion

The aims of our analysis were to investigate the protein-protein interactions of SARS-CoV-2 and to compare a stream of influenza A virus (H1N1) with SARS-CoV-2.

We extracted genes based on their frequencies on the expansion results and then we built a network of the resulting genes. Thanks to the network was possible to restrict the number of genes based on their degree. By choosing a threshold of 10 degree we subselected a list of 213 genes that was used for the enrichment analysis. The enrichment analysis underlined pathways and molecular functions that are well known to be connected with virus-host interaction in literature. In particular, one of the most interesting term, was found to be related with ERAD pathway. COVID-19 replication in infected host cells may perturb protein folding in the endoplasmatic reticulum (ER), inducing a strong ER stress response and continued viral proliferation may induced inflammation and cell death. Moreover, mitochondria term was found enriched in cellular component: viruses, in fact, interfere with mitochondrial pathways to impede mitochondrial anti-vital signalling mechanism.

Based on the dataset available on EnrichR library "Virus_Perturbation_from_GEO_down" and "up" we decided to extract 100 genes due to computational time, based on their importance, and expanded their related isoforms with NES^{2}RA.

Analogously as we did for COVID-19 we built the protein-protein interaction network for H1N1 and we compared it with the one of COVID-19. From this process we extracted 2045 genes and a sub network of 17 genes. We filtered out genes based on their degree to obtain a list of 100 genes that were used for the enrichment analysis.

By applying the PC-algorithm on the sub network of 17 genes and a subsequent drug repurposing, we were able to detect two possible target: PRIM1 and RALA, involved in cell cycle pathways.

From the enrichment analysis of 100 common genes, an interesting result is given by "ribosome biogenesis" term. Ribosome biogenesis factor and ribosomal proteins regulate the translation of viral transcripts and this direct interatctions can suggest a potential new antiviral agents, by targeting ribosome production and function.

By using Laverne Novus Bioinformatics tool with "vascular smooth muscle contraction" as input, the relation with angiotensin II was retrived and was showed to be involved in both COVID-19 and influenza-A (stain H1N1) virus. This a proof of ACE2 vital role in virus-host interactions, since SARS-CoV-2 virus enters in the target cell by binding ACE2, that result in the activation of the renin-angiostensin system, and this make ACE2 a good target.

We have detected possible implication of viruses infection also in gastrointestinal cells, revealing gastrin-releasing peptide, implicated in pulmonary lung inflammatory disease such as influenza, as a potentially novel target for therapeutic intervention also for COVID-19. Additionally, to leverage our analysis with H1N1, we decided to used the list of 100 common genes and search possible matches with H1N1 database on Open target. We had just one match, PLTP, which has no treatment currently available, on Open Target. We exploited also DGIdb and found one FAD approved drug which is linked to cholesterol which is currently studied in order to prevent severe complication of the disease and to reduce the viral ability to infect host cells.

Then we decided to extend our research for new targets to all the Open Target database. It was useless, at this point, to use the 100 common genes. For this reason we decided to use the 213 genes of SARS-CoV-2 used previously. We exploited both Open Target and DGIdb resources and from this emerged some significant and potential targets such as HMOX1, SIRT5 and IMPDH2.

In conclusion our work tried to capture and identify analogy between two viruses that raise from possible zoonosis and started global pandemics H1N1 and SARS-CoV-2. We improved the current understanding of H1N1 virus by expanded 100 genes involved in its infection and extracting a sub network of 17 genes which matched with SARS-CoV-2. Thanks to the PC-algorithm it was possible to observe that, by targeting PRIM1 and RALA we could act on many more genes and so have a wider effect which needs to be further investigated in order to explore the potential side effects that this could have.

References

[1] D. Gordon, G. Jang, M. Bouhaddou, *et al.*, "A sars-cov-2-human protein-protein interaction map reveals drug targets and potential drug-repurposing," *bioRxiv : the preprint server for biology*, 03 2020.

[2] M. Lizio, J. Harshbarger, H. Shimoji, *et al.*, "Gateways to the fantom5 promoter level mammalian expression atlas," *Genome Biology*, vol. 16, p. 22, 01 2015.

[3] M. Lizio, I. Abugessaisa, S. Noguchi, *et al.*, "Update of the fantom web resource: expansion to provide additional transcriptome atlases," *Nucleic acids research*, vol. 47, 11 2018.

[4] E. Y. Chen, C. M. Tan, Y. Kou, *et al.*, "Enrichr: interactive and collaborative html5 gene list enrichment analysis tool," *BMC bioinformatics*, vol. 14, p. 128, April 2013.

[5] M. V. Kuleshov, M. R. Jones, A. D. Rouillard, *et al.*, "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update," *Nucleic Acids Research*, vol. 44, pp. W90–W97, 05 2016.

[6] F. Asnicar, L. Masera, E. Coller, *et al.*, "Nes2ra: network expansion by stratified variable subsetting and ranking aggregation," *The International Journal of High Performance Computing Applications*, vol. 32, no. 3, pp. 380–392, 2018.

[7] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: an open source software for exploring and manipulating networks," in *Third international AAAI conference on weblogs and social media*, 2009.

[8] D. Szklarczyk, A. Gable, D. Lyon, *et al.*, "String v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic acids research*, vol. 47, 11 2018.

[9] D. Ochoa, A. Hercules, M. Carmona, *et al.*, "Open targets platform: supporting systematic drug–target identification and prioritisation, nucleic acids research," *Nucleic Acids Research*, 2020.

[10] K. Cotto, A. Wagner, Y.-Y. Feng, S. Kiwala, A. Coffman, G. Spies, A. Wollam, N. Spies, O. Griffith, and M. Griffith, "Dgidb 3.0: A redesign and expansion of the drug-gene interaction database," *Nucleic acids research*, vol. 46, 11 2017.

[11] S. Payungporn, N. T-Thienprasert, J. Makkoch, and Y. Poovorawan, "Molecular characteristics of the human pandemic influenza a virus (h1n1)," *Acta virologica*, vol. 54, pp. 155–63, 09 2010.

[12] P. Zhou, X. Yang, X.-G. Wang, *et al.*, "A pneumonia outbreak associated with a new coronavirus of probable bat origin," *Nature*, vol. 579, 03 2020.

[13] U. Raudvere, L. Kolberg, I. Kuzmin, *et al.*, "g:profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)," *Nucleic acids research*, vol. 47, 05 2019.

[14] T. Aoe, "Pathological aspects of covid-19 as a conformational disease and the use of pharmacological chaperones as a potential therapeutic strategy," *Frontiers in Pharmacology*, vol. 11, p. 1095, 07 2020.

[15] R. Srivastava, S. Daulatabad, M. Srivastava, and S. C. Janga, "Sars-cov-2 contributes to altering the post-transcriptional regulatory networks across human tissues by sponging rna binding proteins and micro-rnas," *bioRxiv : the preprint server for biology*, 07 2020.

[16] A. West, W. Khoury-Hanold, M. Staron, *et al.*, "Mitochondrial dna stress primes the antiviral innate immune response," *Nature*, vol. 520, pp. 553–557, Apr. 2015.

[17] H. Dong, R. Zhang, Y. Kuang, *et al.*, "Selective regulation in ribosome biogenesis and protein production for efficient viral translation," *Arch Microbiol (2020)*, 2020.

[18] T. Wurm, H. Chen, T. Hodgson, P. Britton, G. Brooks, and J. Hiscox, "Localization to the nucleolus is a common feature of coronavirus nucleoproteins, and the protein may disrupt host cell division," *Journal of virology*, vol. 75, pp. 9345–56, 11 2001.

[19] J. Lamas, M. Alonso-Suarez, J. Fernández Martín, and J. Saavedra-Alonso, "Angiotensin ii suppression in sars-cov-2 infection: a therapeutic approach," *Nefrología (English Edition)*, vol. 40, 06 2020.

[20] L. Chen and G. Hao, "The role of angiotensin-converting enzyme 2 in coronaviruses/influenza viruses and cardiovascular disease," *Cardiovascular Research*, vol. 116, pp. 1932–1936, 04 2020.

[21] P. Zhong, J. Xu, D. Yang, Y. Shen, L. Wang, Y. Feng, C. Du, Y. Song, C. Wu, X. Hu, and Y. Sun, "Covid-19-associated gastrointestinal and liver injury: clinical features and potential mechanisms," *Signal Transduction and Targeted Therapy*, vol. 5, p. 256, 11 2020.

[22] K. Shirey, M. Sunday, W. Lai, M. Patel, J. Blanco, F. Cuttitta, and S. Vogel, "Novel role of gastric releasing peptide-mediated signaling in the host response to influenza infection," *Mucosal Immunology*, vol. 12, p. 1, 10 2018.

[23] M. P. A. S. M. B. Dina Radenkovic, Shreya Chawla, "Cholesterol in relation to covid-19: Should we care about it?," *Journal of Clinical Medicine*, vol. 9, 06 2020.

[24] R. El Baba and G. Herbein, "Management of epigenomic networks entailed in coronavirus infections and covid-19," *Clin Epigenet*, vol. 27, 2020.

[25] D. Singh, H. Wasan, and K. Reeta, "Heme oxygenase-1 modulation: A potential therapeutic target for covid-19 and associated complications," *Free Radical Biology and Medicine*, vol. 161, 2020.

[26] N. Bukreyeva, E. Mantlo, R. Sattler, C. Huang, S. Paessler, and J. Zeldis, "The impdh inhibitor merimepodib suppresses sars-cov-2 replication in vitro," 04 2020.

[27] W. Lee, L. Chen, C. Lee, *et al.*, "Dna primase polypeptide 1 (prim1) involves in estrogen-induced breast cancer formation through activation of the g2/m cell cycle checkpoint," *International Journal of Cancer*, vol. 144, 08 2018.

All the script and materials that have been used and created to perform these analysis can be found in an online repository.[1]

[1] https://github.com/FrancescoPenasa/UNITN-CS-2020-LabOfBiologicDataMining