Journal of proteome research

pubs.acs.org/jpr

Article

Computational Identification of Human Biological Processes and Protein Sequence Motifs Putatively Targeted by SARS-CoV-2 Proteins Using Protein−Protein Interaction Networks

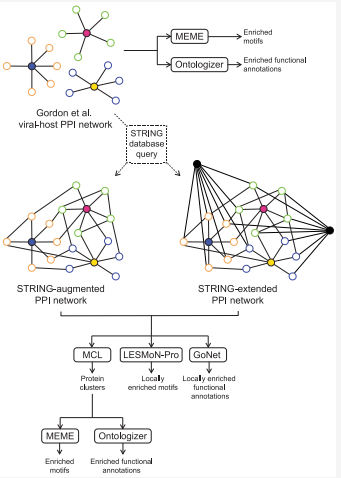
Rachel Nadeau,[§] Soroush Shahryari Fard,[§] Amit Scheer,[§] Emily Hashimoto-Roth,[§] Dallas Nygard,[§] Iryna Abramchuk,[§] Yun-En Chung,[§] Steffany A. L. Bennett, and Mathieu Lavallée-Adam*

Cite This: *J. Proteome Res.* 2020, 19, 4553−4566

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: While the COVID-19 pandemic is causing important loss of life, knowledge of the effects of the causative SARS-CoV-2 virus on human cells is currently limited. Investigating protein−protein interactions (PPIs) between viral and host proteins can provide a better understanding of the mechanisms exploited by the virus and enable the identification of potential drug targets. We therefore performed an in-depth computational analysis of the interactome of SARS-CoV-2 and human proteins in infected HEK 293 cells published by Gordon et al. (*Nature* **2020**, *583*, 459−468) to reveal processes that are potentially affected by the virus and putative protein binding sites. Specifically, we performed a set of network-based functional and sequence motif enrichment analyses on SARS-CoV-2-interacting human proteins and on PPI networks generated by supplementing viral-host PPIs with known interactions. Using a novel implementation of our GoNet algorithm, we identified 329 Gene Ontology terms for which the SARS-CoV-2-interacting human proteins are significantly clustered in PPI networks. Furthermore, we present a novel protein sequence motif discovery approach, LESMoN-Pro, that identified 9 amino acid motifs for which the associated proteins are clustered in PPI networks. Together, these results provide insights into the processes and sequence motifs that are putatively implicated in SARS-CoV-2 infection and could lead to potential therapeutic targets.



KEYWORDS: *COVID-19, SARS-CoV-2, protein−protein interaction network, motif discovery, gene ontology, enrichment analysis, clustering, statistics, graph theory*

INTRODUCTION

The COVID-19 (Coronavirus Disease 2019) pandemic is causing massive loss of life around the globe and has had a dramatic impact on healthcare systems and economies worldwide. The virus at the root of this pandemic, SARS-CoV-2, is a highly pathogenic coronavirus[1] that spreads with great efficiency.[2] While vaccines are currently in development to contain the pandemic, the efficacy of those vaccines is still undetermined,[3] and their worldwide availability will take time. Furthermore, while the therapeutic agent Remdesivir has been shown to the reduce recovery time of hospitalized COVID-19 patients, no drugs are currently approved for treatment.[4] Adding efficacious drugs to our arsenal to fight SARS-CoV-2 infections would strengthen our ability to dampen the impacts of the pandemic. Insights into the host processes that are targeted during SARS-CoV-2 infection would improve our drug development capabilities and potentially suggest drugs, for which safety has already been determined, that could be repurposed for use against SARS-CoV-2. It is therefore critical to derive a better understanding of the mechanisms by which SARS-CoV-2 infects and causes disease in human host cells. Furthermore, SARS-CoV-2 constitutes the third highly pathogenic coronavirus, which has presented a serious threat to the globe, with SARS-CoV and MERS-CoV also having caused significant loss of life in the past.[5,6] Nevertheless, very little is known about their respective mechanisms, nor do drugs exist to treat their infections. The recurrence of pathogenic coronavirus outbreaks suggests a significant likelihood of yet another future pathogenic coronavirus emergence. Hence, any understanding we gain on the SARS-CoV-2 infection and identification of compounds that are effective for use as a treatment could play a critical role in containing future pathogenic coronavirus outbreaks.

Protein−protein interactions (PPIs) are extremely useful to map out biological processes, protein machineries, and protein complexes.[7] In an effort to provide a better understanding of the biological processes affected by SARS-CoV-2 in human

Special Issue: Proteomics in Pandemic Disease

Received: June 15, 2020
Published: October 26, 2020

https://dx.doi.org/10.1021/acs.jproteome.0c00422
J. Proteome Res. 2020, 19, 4553−4566

host cells, Gordon et al.[18] mapped the interactions of SARS-CoV-2 proteins with human proteins in infected HEK 293 cells. Their effort early on during the outbreak highlighted several potential drug targets and drug candidates. It also generated an interactome containing a great wealth of information that can be further analyzed to identify novel insights into SARS-CoV-2 mechanisms of infection and host−pathogen interactions.

Algorithms including the Markov Clustering algorithm (MCL),[8] Restricted Neighborhood Search Clustering Algorithm (RNSC),[7] MCODE,[9] Socio-Affinity Index,[10] and the eigenmode analysis of the connectivity matrix of a PPI network[11] have proven effective at characterizing protein complexes in PPI networks using clustering strategies. Such clusters can then be investigated for functional enrichment by determining whether a given function is overrepresented among the proteins within the cluster with respect to the number of proteins it annotates in the entire network. Gene Ontology terms,[12] KEGG pathways,[13] and REACTOME pathways[14] are annotations that are often used for such enrichment analyses.

We have shown in the past with our tool GoNet that evaluating the clustering of GO terms in PPI networks can be done to identify biological processes and protein complexes of interest.[15] We have also shown that PPI networks can be effectively used to discover novel RNA sequence motifs that are associated with groups of significantly clustered proteins using another algorithm called LESMoN.[16] Herein, we propose a set of functional enrichment analyses and protein sequence motif discovery approaches to thoroughly investigate the set of human proteins interacting with SARS-CoV-2 proteins revealed by Gordon et al.[18] We first directly analyze their network of PPIs to identify GO terms, cellular localizations, and protein sequence motifs that are enriched among SARS-CoV-2-interacting human proteins. We then supplement this set of interactions with known human PPIs to create additional networks and apply clustering approaches to identify functional annotations and protein sequence motifs that are clustered in these networks. Given the relevance of PPIs in defining the effects of viral proteins on host processes, these functional annotations and sequence motifs are likely to reveal previously unappreciated aspects of SARS-CoV-2 infection. Our analyses provide a better understanding about the effect of SARS-CoV-2 on host cell machinery and have the potential to help in the discovery or repurposing of drugs for COVID-19.

EXPERIMENTAL SECTION

Overview

We present a suite of enrichment analyses investigating both functional and sequence motif overrepresentation (Figure 1). We directly analyze the viral-host PPI network generated by Gordon et al.[18] and also build from it an augmented and an extended network using human PPIs from the STRING database.[17] We perform functional and sequence motif enrichment analyses on these PPI networks. We identify GO terms, cellular locations, and sequence motifs that are clustered in the human PPI networks, derived with the help of STRING, using different clustering approaches.

Data Set

Our computational analyses were performed on three PPI networks. The first one is the high-confidence viral-host PPI network of SARS-CoV-2 proteins interacting with *H. sapiens*

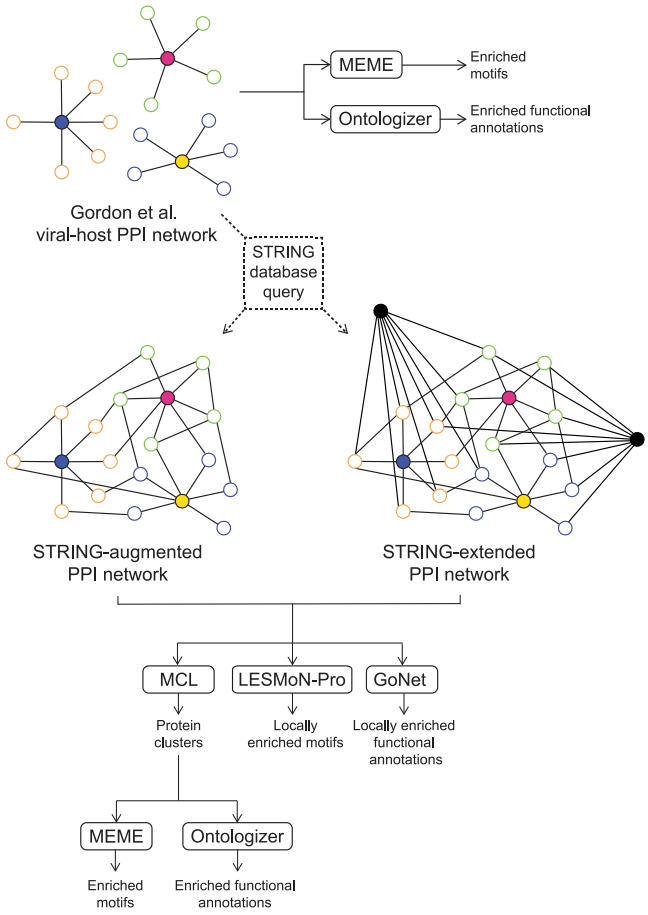


Figure 1. Graphical representations of the PPI networks analyzed and the enrichment analysis approaches applied on them. Color-filled proteins in the PPI networks represent SARS-CoV-2 proteins. Proteins without color filling represent *H. sapiens* proteins interacting with SARS-CoV-2 proteins. Black-filled proteins represent *H. sapiens* superinteractors.

HEK 293 proteins generated by Gordon et al.[18] Since this network was built solely by the affinity purification of SARS-CoV-2 proteins, the interactions between *H. sapiens* proteins are not included in this network. We therefore built a second version of this network by adding to it interactions between human proteins in the viral-host PPI network based on the information stored in the STRING PPI database (downloaded on April 2, 2020).[17] Human proteins that were purified by SARS-CoV-2 proteins were therefore connected in this network if they reached a medium confidence (0.4) according to STRING and had either an experimental or database evidence. This resulted in the creation of a network containing 216 proteins and 502 PPIs. However, not all of these proteins are connected to each other. We therefore identified the largest connected component of the network and discarded the remaining proteins, which resulted in a network of 195 proteins and 489 PPIs (Supplementary Table S1). This network will be referred to as the STRING-augmented network. In order to further analyze the biological processes that may be affected by SARS-CoV-2 proteins, we also built an extended network (third version), which includes proteins that are tightly interacting with *H. sapiens* interactors of SARS-CoV-2 proteins. Such proteins could possibly be affected through downstream effects of viral-host protein−protein interactions or even may have been missed in the original affinity

https://dx.doi.org/10.1021/acs.jproteome.0c00422
J. Proteome Res. 2020, 19, 4553−4566

purification experiments. These proteins were identified in STRING as *H. sapiens* proteins that are interacting with at least eight *H. sapiens* proteins that are themselves interacting with SARS-CoV-2 proteins. Eight was chosen as the minimum number of interactors to limit the size of the network and maintain a high degree of connectivity between proteins. We have named these proteins superinteractors. A STRING interaction was counted for those superinteractors if it achieved a medium confidence score (0.4) and had either an experimental or database evidence. This strategy ensured that only proteins with a close relationship with SARS-CoV-2 protein interactors would be included in the network. The addition of those proteins resulted in the creation of a network made of a single connected component of 867 proteins and 32 549 PPIs (Supplementary Table S2). This network will be referred to as the STRING-extended network. All three networks are graphically represented with toy examples in Figure 1 along with the computational methods applied to them, which are described below.

Markov Clustering Algorithm to Identify Clusters in the STRING-Augmented and -Extended Networks

We used the Markov Clustering Algorithm (MCL)[8] to identify clusters of proteins in the STRING-augmented and -extended networks. The algorithm was executed using an inflation parameter of 2 to obtain a reasonable level of granularity in the protein clusters.

Functional Enrichment Analysis

In order to identify overrepresented biological processes, cellular components, and molecular functions in the viral-host PPI network and the clusters obtained by the MCL analyses of the STRING-augmented and -extended networks, a GO enrichment analysis was performed using the Ontologizer package.[19] An analysis was performed for each individual set of proteins purified by a SARS-CoV-2 protein, the set of those proteins along with their superinteractors (proteins interacting with 8 or more proteins detected by Gordon et al.[18]), and for each cluster of size 3 or greater, as detected by MCL in the STRING-augmented and -extended networks. The set of all proteins present in the respective networks was used as background for the enrichment analyses. Ontologizer uses a modified Fisher's exact test to assess the statistical significance of a GO term enrichment. The resulting *p*-values were adjusted for multiple hypothesis testing using the Benjamini−Hochberg procedure[20] with Ontologizer. Similarly, we also investigated the same subnetworks for the enrichment of annotations related to cellular location, as derived from the Cell Map.[21] A Fisher's exact test was again used to assess statistical significance of the enrichments and the resulting *p*-values were adjusted for multiple hypothesis with the Benjamini−Hochberg procedure.

Protein Sequence Motif Enrichment Analysis

With the objective of discovering protein sequence motifs that are surprisingly overrepresented in certain sections of the networks, we performed a motif enrichment analysis using the MEME Suite[22] on network elements of the viral-host network and the STRING-augmented and -extended networks. As for the GO enrichment analysis, MEME was executed on each individual set of proteins purified by a SARS-CoV-2 protein, as well as those sets supplemented with their superinteractors. It was also executed on all clusters of size 3 or more, as detected by the MCL algorithm, in both STRING-augmented and -extended networks. MEME was executed in classic mode with a 0-order Markov Model and zero or one occurrences per site for motifs with a minimum width of 5 and a maximum width of 30.

Identifying GO Terms That Are Clustered in the STRING-Augmented and -Extended Networks Using GoNet 2.0

To investigate the clustering of proteins sharing the same GO terms in the networks, we implemented a modified version of our previously published approach named GoNet.[15]

Clustering Significance Assessment. Briefly, as previously described,[15] GoNet measures the clustering of a GO term *t*, annotating *P* proteins by calculating its total pairwise distance (TPD) in the network, which in this context, is the sum of the shortest paths between all pairs of proteins annotated with *t*. GoNet then assesses the significance of this clustering measure using a Monte Carlo sampling approach. We modified GoNet's statistical assessment approach to take into account the annotation bias of proteins, where proteins that have many protein interactions tend to be more heavily annotated than proteins with fewer interactions and vice versa. To assess the significance of the total pairwise shortest path of *t*, GoNet randomly samples without replacement *P* proteins in the network 100 000 times and computes the total pairwise distance between all proteins in each sample. The difference in this new version of GoNet is that the sampling probability of each protein is roughly proportional to the number of GO terms annotating it. This procedure ensures that proteins annotated by a large number of GO terms are sampled more often than those with fewer GO terms, since the former proteins will see their clustering tested more often in the network. Using an approach inspired by our clustering significance assessment tool, LESMoN,[16] we estimated the mean and standard deviation of the TPD for each value of *P* and derived a normal distribution of the TPD for all *P*'s. These distributions can be used as null models to estimate *p*-values for the TPDs of all GO terms. GO terms annotating three proteins or more in the PPI networks saw their clustering significance assessed. The clustering of proteins associated with a total of 1853 and 4393 GO terms was evaluated in the STRING-augmented and -extended networks, respectively.

False Discovery Rate Estimation. Since GO terms share several protein annotations, the statistical testing of GO term clustering is not independent. Traditional multiple hypothesis testing correction strategies, such as the Bonferroni correction, are therefore likely to be overly conservative in their *p*-value adjustments. Instead, again inspired by our LESMoN algorithm, we used a permutation-based strategy to estimate false discoveries. Briefly, pairs of GO-protein annotations are randomly selected and swapped, such that if protein A is annotated by GO term t_1 and protein B is annotated by GO term t_2, after their random selection and swapping, A will be annotated by t_2 and B by t_1. This annotation swapping procedure is performed 1000× the total number of GO-protein associations times, such that the annotation-permuted network is representative of a randomly annotated network. After this annotation permutation, GoNet is then executed, such that the statistical significance of the clustering of the permuted GO terms is assessed. The false discovery rate (FDR) at a given *p*-value α threshold is then estimated as follows:

https://dx.doi.org/10.1021/acs.jproteome.0c00422
J. Proteome Res. 2020, 19, 4553−4566

$$\mathrm{FDR}(\alpha) = \min\left(\mathrm{FDR}(\alpha + \mathrm{inc}), \left(\frac{\frac{\#\text{ of permutated GO terms with } p\text{-value} < \alpha}{\#\text{ of permuted GO terms}}}{\frac{\#\text{ of GO terms with } p\text{-value} < \alpha}{\#\text{ of GO terms}}}\right)\right)$$

where inc is a small increment in p-values. FDR of α corresponds to the minimum between the FDR estimated for a slightly higher p-value and the FDR at α, such that the FDR function remains monotonic, preventing noisy FDR fluctuations at very small p-value thresholds.

Identifying Protein Sequence Motifs That Are Locally Enriched in the STRING-Augmented and -Extended Networks Using LESMoN-Pro

We have previously presented an approach called LESMoN, which identifies 5′ untranslated sequence motifs, for which the associated proteins are clustered in a PPI network.[16] Herein, we propose a new version of LESMoN, called LESMoN-Pro, which detects protein sequence motifs for which the associated proteins are clustered in a PPI network. LESMoN-Pro uses similar algorithmic principles as LESMoN, while being adapted to search for protein motifs instead of RNA motifs. Briefly, LESMoN-Pro enumerates amino acid motifs of size 8 over the following alphabet: A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y, X, B, Z, J, l, @, h, o, p, t, s, b, +, −, c, wherein the following are degenerate characters encompassing multiple amino acids: B (asparagine and aspartic acid), Z (glutamine and glutamic acid), J (leucine and isoleucine), l (aliphatic), @ (aromatic), h (hydrophobic), o (alcohol), p (polar), t (tiny), s (small), b (bulky), + (positively charged), − (negatively charged), c (charged), and X is a wild card character corresponding to all amino acids. All other characters correspond to single amino acids. The exact sets of amino acids mapping to the different degenerate characters are given in Supplementary Table S3. Using this alphabet, LESMoN-Pro enumerates motifs that are represented in the set of protein sequences of the largest connected component of both the STRING-augmented and -extended PPI networks, such that a maximum of 4 degenerate characters can appear in any given motif. This rule, along with the maximum motif length of 8, ensures that the motifs are not too degenerate (i.e., present in most, if not all, proteins) and also keeps the algorithm's running time reasonable. Protein sequences were downloaded from the UniProt SwissProt database[23] on Feb. 12, 2020. Motifs that matched to the sequences of at least 3 proteins in the networks were retained for downstream analyses. The clustering of the proteins associated with a total of 48 128 026 motifs was evaluated in the STRING-augmented network and 514 670 598 motifs in the STRING-extended network. Motifs were associated with sets of proteins ranging in size from 3 to 34 proteins in the STRING-augmented network and 3 to 99 in the STRING extended network.

Clustering Significance Assessment. LESMoN-Pro's clustering significance assessment of motifs in the PPI networks mirrors that of GoNet. Indeed, one could consider that a sequence motif contained in a protein sequence is annotating that protein, much like a GO term annotates a protein. Hence, the clustering of proteins containing a given motif was measured using the total pairwise distance given by the shortest paths between all pairs of proteins. A Monte Carlo sampling approach was then used to assess the statistical significance of the clustering. Once again, since some proteins are likely to contain more motifs than others due to their sequence length or amino acid composition, sampling of individual proteins is performed with a probability roughly proportional to the number of motifs associated with a given protein. For all sizes of sets of proteins annotated by a motif, a null distribution of the TPD is estimated using a normal distribution approximated using the mean and standard deviation obtained from 100 000 random samplings of protein sets. A p-value is then estimated by comparing the TPD of a given motif to its corresponding null distribution.

False Discovery Rate Estimation. In order to estimate an FDR at a given p-value threshold, protein sequences were locally shuffled, as previously described.[16] Briefly, amino acids within nonoverlapping sliding windows of size 10 along all protein sequences were shuffled. 10 000 pairs of amino acids were randomly selected within each window and permuted. This procedure ensures that any local sequence properties are maintained, while generating sequences that are not likely to be present in the human proteome. The LESMoN-Pro assessment of clustering significance is then applied on the motifs present in those locally shuffled sequences. Using the same strategy as GoNet, an FDR is estimated for a given p-value.

Determining Sequence Motif Families of Biological Interest. Since a number of interacting proteins in the STRING-augmented and -extended networks share important levels of sequence homology, a number of motifs found are likely to simply originate from highly homologous regions. While we acknowledge that such motifs are not devoid of biological interest, we claim that they are not as interesting, or at least as surprising, as motifs conserved in interacting proteins that do not share a high level of sequence homology. We therefore built a procedure for filtering out motifs that are mainly contained in highly homologous protein sequences. Specifically, protein sequence identity between all pairs of proteins was computed by Clustal Omega.[24] Motifs for which 75% of the pairs of the associated proteins showed a sequence identity >25% were deemed to be originating primarily from homologous sequences and discarded. Furthermore, similar motifs are likely to share a similar level of clustering significance since they are likely to be associated with a similar set of proteins. In order to avoid reporting redundant results, we used LESMoN's approach for motif family grouping.[16] Specifically, a hierarchical clustering approach using average-linkage and a similarity measure based on the fraction of shared proteins between pairs of motifs was used to identify families of similar motifs. Motif families are determined by performing a cut in the resulting dendrogram. A representative motif for each family is then determined using the motif with the best clustering p-value. In the event of a tie, the motif that is the least substituted with degenerate characters, based on the number and specificity of the degenerate characters, is selected. If no single motif is identified as the least degenerate among the motifs with the lowest p-value, one motif from the list of such motifs is randomly selected as the representative motif. Finally, a consensus motif is then built using ggseqlogo[25] based on the matching occurrences of the representative motifs in the associated protein sequences.

https://dx.doi.org/10.1021/acs.jproteome.0c00422
J. Proteome Res. 2020, 19, 4553−4566

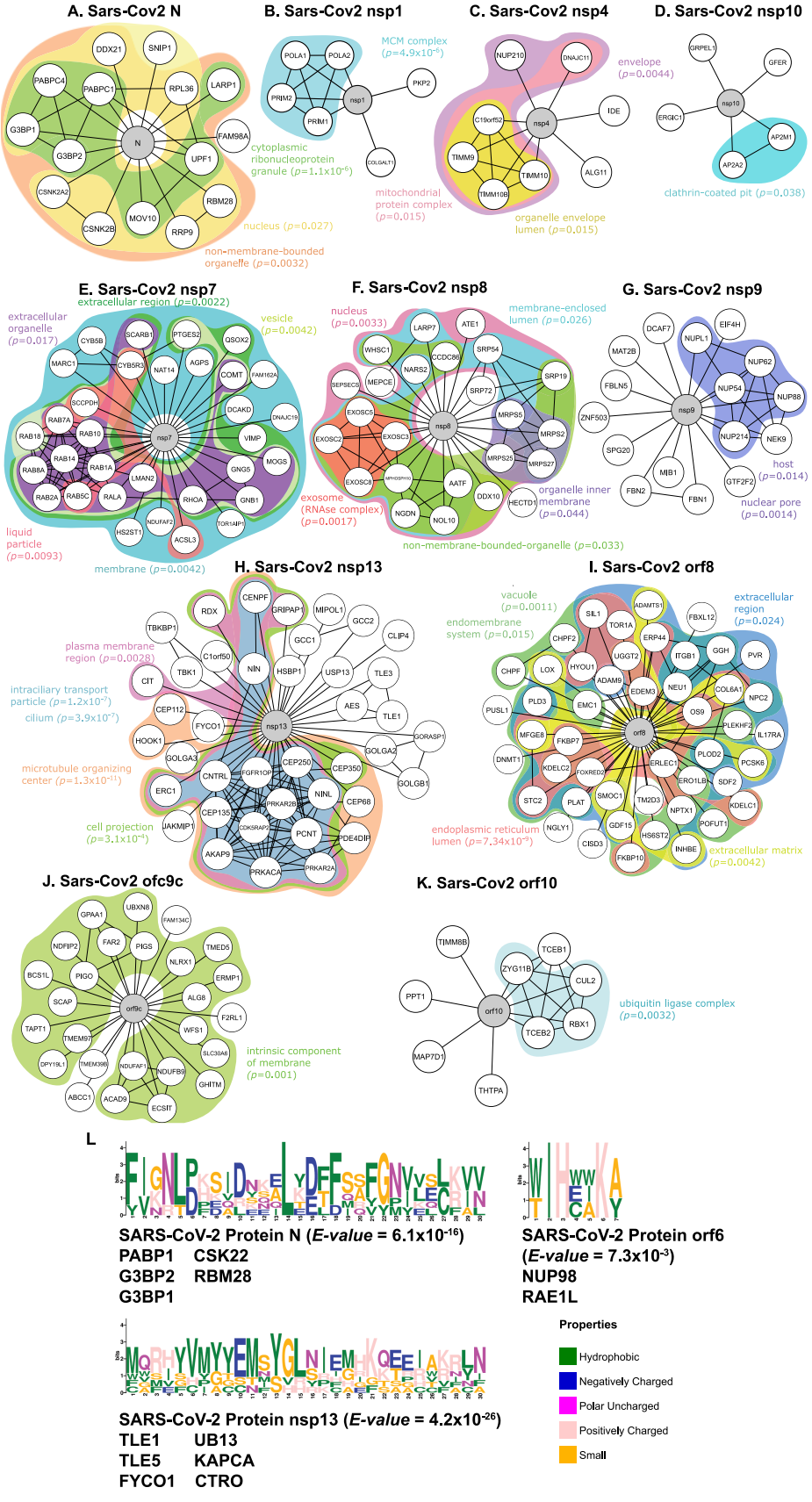


Figure 2. Enrichment analyses on *H. sapiens* interactors of a selected set of SARS-CoV-2 proteins. (A–K) GO enrichment analysis of protein interactors of SARS-CoV-2 proteins. GO cellular components are color-coded. *p* corresponds to an FDR-adjusted *p*-value. (L) Motifs enriched among the protein sequences of the *H. sapiens* interactors of SARS-CoV-2 proteins N, nsp13 and orf6. Proteins containing the motifs are listed below the motif logos. Amino acids are color-coded based on their properties.

https://dx.doi.org/10.1021/acs.jproteome.0c00422
J. Proteome Res. 2020, 19, 4553–4566

Motif Location in Protein Domains

The location of the family representative motifs was mapped onto known protein family domains of each protein within which they appeared based on the sequence information from UniProt[23] and domain annotations from the Pfam[26] protein family domain database.

Software Availability

All software packages developed and implemented for GoNet and LESMoN-Pro are open-source and accessible at this address: https://github.com/LavalleeAdamLab/sars-cov-2.

RESULTS

We present an in-depth enrichment analysis of PPI networks involving interactions between SARS-CoV-2 and *H. sapiens* proteins. We apply functional and motif enrichment analyses on sets of proteins that were purified by SARS-CoV-2 proteins in the Gordon et al.[18] study. We also generate PPI networks of human proteins that are interacting with SARS-CoV-2 proteins, based on interactions derived from the STRING database. Using these networks, we use clustering strategies and enrichment analysis tools to reveal functional annotations and protein sequence motifs for which the associated proteins are significantly clustered in the PPI networks.

Human Proteins Purified by SARS-CoV-2 Proteins Show Functional Enrichments

While Gordon et al.[18] performed a GO enrichment analysis on the proteins they have purified using SARS-CoV-2 proteins, they only performed this analysis for biological processes. We complemented this analysis investigating both GO cellular components (Figure 2A–K) and molecular functions (Supplementary Figure S1), in addition to having repeated the analysis for biological processes using a different algorithm (i.e., Ontologizer) (Supplementary Figure S2), with the goal of providing a better understanding of the functional role played by SARS-CoV-2 proteins in infection. Complete Ontologizer GO enrichment analysis results are provided in Supplementary File S1. These enrichment analyses highlighted, among other findings, that the SARS-CoV-2 protein N interactors were enriched for the term "Nucleus" (FDR-adjusted p-value (p) = 0.027), the PPIs of nsp1 were enriched for "MCM complex" proteins ($p = 4.9 \times 10^{-6}$), and nsp8 interactors were overrepresented by "nucleus" proteins ($p = 0.033$), while the interactors of nsp9 were enriched for "nuclear pore" proteins ($p = 0.0014$), suggesting that these proteins may play a role in the nucleus of the host cell (Figure 2). On the other hand, SARS-CoV-2 nsp7 show enrichment among its interactors for "membrane" ($p = 0.0042$) and "vesicle" proteins ($p = 0.042$). Interestingly, SARS-CoV-2 nsp13's PPIs were highly enriched for microtubule organizing center proteins ($p = 1.3 \times 10^{-11}$), while orf10's interactors were enriched for ubiquitin ligase complex proteins ($p = 0.0032$).

In addition to Gene Ontology annotations, we analyzed whether SARS-CoV-2 protein interactors were enriched for specific cellular locations, as defined by the Cell Map[21] annotations (Supplementary Figure S3; Supplementary File S2). Such annotations are complementary to those related to cellular components in Gene Ontology and were derived from a network of interactions mapped using a proximity labeling approach.[21] Consistent with the GO enrichment analysis, the Cell Map location enrichment analysis also saw an enrichment of membrane-related proteins among the interactors of nsp7 ($p = 3.4 \times 10^{-6}$). In addition, this analysis revealed the enrichment of "ER membrane" ($p = 0.00011$) and "mitochondrial matrix" proteins ($p = 0.0023$) among the interactors of the SARS-CoV-2 M protein. It also highlighted that the interactors of the viral nsp13 protein are enriched for the "centrosome" ($p = 0.00020$) and "microtubule cytoskeleton" ($p < 10^{-8}$) annotations and those of orf8 are enriched for "ER lumen" ($p = 3.8 \times 10^{-8}$).

We also performed these functional enrichment analyses on the set of *H. sapiens* proteins interacting with SARS-CoV-2 proteins and their superinteractors (see Experimental Section; Supplementary File S3). This analysis revealed biological processes, such as "gene silencing", "membrane docking", and "nuclear transport" that are enriched among these sets of proteins (Supplementary Figure S4; $p < 0.01$). It also highlights that cellular components related to the vacuole, envelope, endomembrane and microtubule are enriched among such interactors of SARS-CoV-2 proteins ($p < 0.01$). Moreover, molecular functions related to ubiquitination, actin binding, chromatin binding and transporter activity were found to be enriched among these SARS-CoV-2 protein interactors and superinteractors (Supplementary Figure S4; $p < 0.01$). Finally, Cell Map location annotations were also analyzed for enrichment among these proteins. The addition of the superinteractors revealed additional annotations related to, among other localizations, the lysosome, plasma membrane, and chromatin (Supplementary Figure S5; Supplementary File S4; $p < 0.05$).

Human Proteins Purified by SARS-CoV-2 Proteins Are Enriched for Specific Motifs

In addition to functional enrichment, we investigated whether the groups of proteins interacting with SARS-CoV-2 proteins were enriched for specific amino acid sequence motifs using the MEME Suite[22] (Supplementary File S5). Such motifs could play an important role in the binding of SARS-CoV-2 proteins or in the host mechanism these proteins are affecting. A number of SARS-CoV-2 proteins showed such enrichments (E-value <0.05). Among these, SARS-CoV-2 proteins N, nsp13, and orf6 interacted with proteins sharing similar sequence motifs (Figure 2L). It is worth noting that many more motifs were identified, but a large portion of them, such as those identified for nsp7, appeared to be detected by MEME simply because the interactors of this protein, RAB-family proteins, share a high level of homology (Supplementary File S5).

We also investigated sequence motif enrichments among the set of SARS-CoV-2 protein interactors and their superinteractors. This analysis could help reveal motifs or functional domains that are affected by processes downstream of those directly mediated by SARS-CoV-2. Our analysis revealed motifs that did not appear to be the sole result of high protein sequence homology, and that were enriched among such interactors for proteins E, M, and N (Supplementary Figure S6; E-value < 0.05). Full results are provided in Supplementary File S6.

Clustering Analysis Reveals Biological Processes That Are Putatively Affected by SARS-CoV-2

In an effort to characterize the biological processes, cellular components, and molecular functions that are affected by SARS-CoV-2 infection, we supplemented the viral-host PPI network of Gordon et al.[18] with STRING PPIs in order to connect the human proteins based on known interactions (STRING-augmented network). This enabled us to perform a

https://dx.doi.org/10.1021/acs.jproteome.0c00422
J. Proteome Res. 2020, 19, 4553−4566

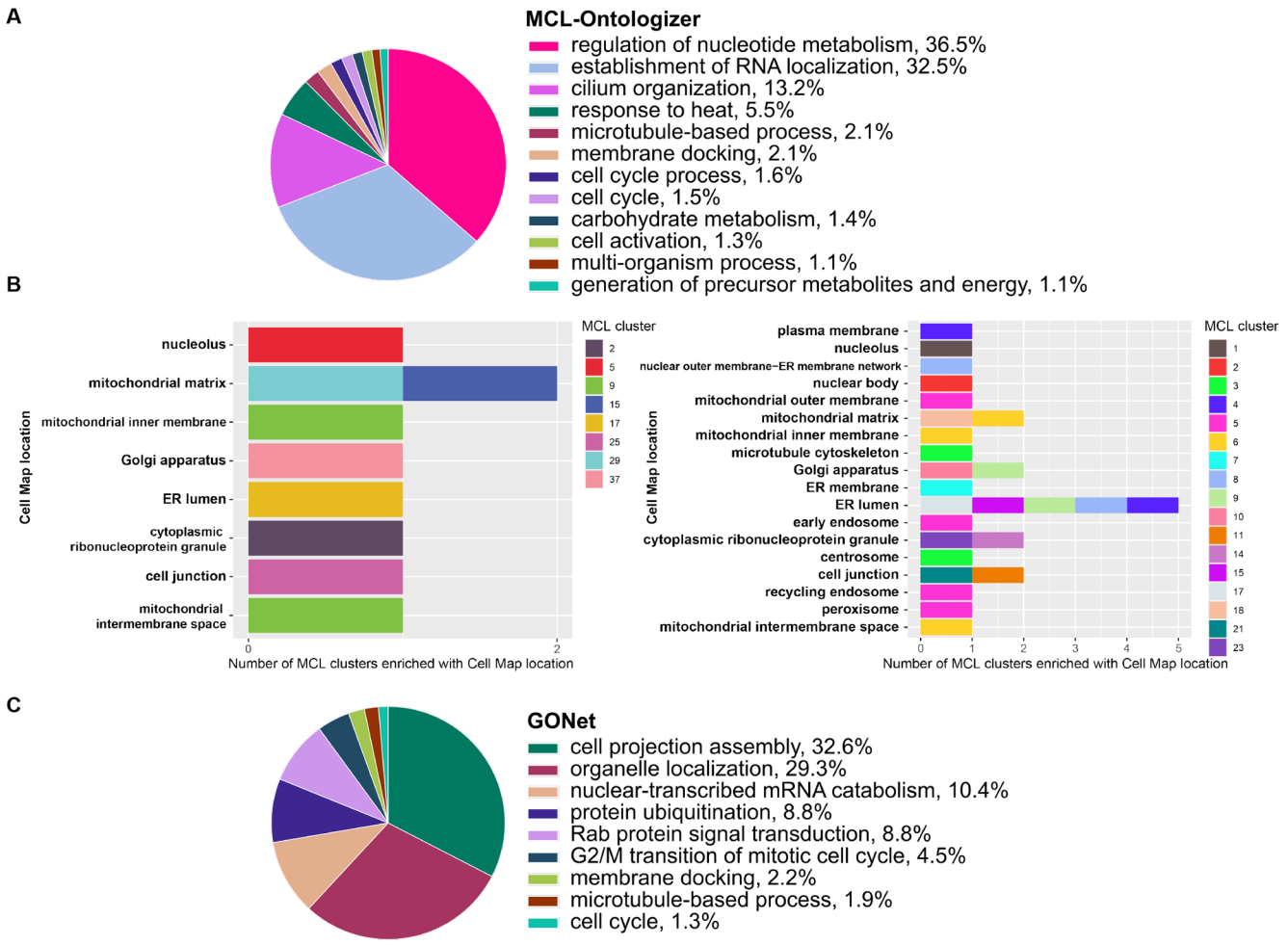


Figure 3. Clustering of functional annotations in the PPI networks. (A) Pie charts of the GO enrichment analysis of MCL clusters from the STRING-augmented network (FDR-adjusted *p*-value <0.01). When a GO biological process was enriched in more than one MCL clusters, the lowest FDR-adjusted *p*-value was used to generate the pie chart. (B) Cell Map location enrichments in MCL cluster from the STRING-augmented (left) and -extended networks (right) (FDR-adjusted *p*-value <0.05). (C) Clustering statistical significance of GO biological processes according to GoNet in the STRING-augmented network (FDR < 0.01). (A,C) GO biological processes with the highest level of enrichment statistical significance are attributed larger pieces of the pie. The portion occupied by each GO term is also represented in percentages next to the term names. Significant GO annotations were processed with REVIGO[56] to summarize the main annotations and remove redundancy. REVIGO output was then fed as input to CirGO[57] for pie chart visualization.

clustering analysis, using the MCL algorithm, in order to identify *H. sapiens* proteins that are interacting with SARS-CoV-2 proteins and also closely interacting together. Our MCL analysis revealed 29 clusters of at least 3 proteins (Supplementary Table S4). A GO enrichment analysis on these clusters revealed several GO biological processes that were significantly overrepresented (FDR-adjusted *p*-value (p) < 0.01; Figure 3A). Among the main GO biological processes enriched in these clusters, we find GO terms related to the establishment of RNA localization, metabolism, membrane docking, cell cycle, among other GO terms. We also find GO cellular components related to nuclear pores, microtubules, cilia, and cell projections to be enriched (p < 0.01; Supplementary Figure S7A). Finally, GO molecular functions such as transporter activity, lyase activity, and structural molecule activity are enriched in the clusters (p < 0.01; Supplementary Figure S8A). Complete GO enrichment analysis results of the MCL clusters are available in Supplementary File S7. These GO terms highlight the main functional annotations that are likely affected by SARS-CoV-2.

The same analysis was performed on the STRING-extended network, yielding 19 clusters of at least 3 proteins (Supplementary Table S5; Supplementary File S8). Even though some of the clusters differed, since the networks were different, the GO terms enriched in those clusters were generally similar to those found in the STRING-augmented network (Supplementary Figure S9; p < 0.01). Indeed, GO biological processes such as "membrane docking" and terms related to the cell cycle were again found to be clustered (p < 0.01). However, there were some notable additions, such as "mRNA processing" and "immune system process", which were found to be enriched in MCL clusters (Supplementary Figure S9A; p < 0.01). In terms of cellular components, the STRING-extended network saw more enrichments among its MCL clusters (Supplementary Figure S9B). Indeed, GO terms related to the vacuole, vesicle, intraciliary transport particle, and respiratory chain were all found to be significantly enriched in the clusters of the STRING-extended network (p < 0.01). The same applies for molecular function GO terms, which saw a number of additional clustered terms, such as

https://dx.doi.org/10.1021/acs.jproteome.0c00422
J. Proteome Res. 2020, 19, 4553−4566

ubiquitin-like related function (Supplementary Figure S9C; $p < 0.01$).

Interestingly, when investigating the MCL clusters for enrichment using the Cell Map location annotations, our analysis revealed a number of supplementary annotations that were significantly clustered within the MCL clusters from both STRING-augmented and -extended networks (Figure 3B; FDR-adjusted p-value <0.05). Those included mitochondrial annotations, as well as cell junction and Golgi apparatus. Overall, the STRING-extended network revealed more annotations that were enriched among its MCL clusters. On top of those, annotations related to the endosome, centrosome, and peroxisome were highlighted by our analysis (Figure 3B; FDR-adjusted p-value <0.05).

To further investigate the clustering of GO terms in the STRING-augmented network, we propose a novel adapted version of our GoNet algorithm, which detects GO terms for which the proteins are surprisingly clustered in a PPI network (Supplementary Table S6). The advantage of this approach over MCL is that no clusters need to be predetermined before GO enrichment analysis. Indeed, MCL could break apart a large cluster into two smaller clusters and therefore separate a given GO term, which could be enriched in the large cluster. GoNet instead simply evaluates the clustering of GO terms and is therefore not affected by overlapping clusters. Overall, GoNet identified 329 GO terms for which the associated proteins were significantly clustered in the STRING-augmented PPI network (p-value <0.001 and FDR < 0.01). GoNet detected GO biological processes that MCL highlighted, but also uniquely identified GO terms related to protein ubiquitination and RAB protein signal transduction, which were found to be significantly clustered (FDR < 0.01; Figure 3C). GoNet also identified GO cellular components related to centrosomes, cell division sites, and vesicles that the MCL-based approach did not highlight (FDR < 0.01; Supplementary Figure S7B). Finally, GoNet found additional GO molecular functions related to myosin binding, protein kinase A binding, and repressing transcription factor binding to be locally enriched in the STRING-augmented network, which MCL did not highlight (FDR < 0.01; Supplementary Figure S8B). Once again, these significantly clustered GO terms highlight groups of highly interacting proteins involved in similar functions that are known to be bound, and therefore putatively affected, by SARS-CoV-2. We also applied our GoNet algorithm to identify GO terms that are clustered in the STRING-extended network and detected 1109 GO terms that were significantly clustered (FDR < 0.01; Supplementary Table S7). The results were generally very similar to those of the STRING-augmented network, with biological processes such as "membrane docking", "nuclear-transcribed mRNA catabolism", "Rab protein signal transduction", and those related to the cell cycle found as significantly clustered in the network (FDR < 0.01; Supplementary Figure S10A). As for cellular components, the results were also generally similar; with the addition of the GO term "vacuole" that was found to be significantly clustered in the STRING-extended network (FDR < 0.01; Supplementary Figure S10B), supporting the results from our previous MCL analysis. Finally, several similarities were found in terms of molecular functions, with the notable addition of ubiquitin-related functions that were detected to be clustered in the STRING-extended network (FDR < 0.01; Supplementary Figure S10C), again supporting the results of the MCL analysis.

Protein Annotated with Gene Ontology Terms Identified as Clustered by GoNet in the STRING-Extended Network Are Differentially Expressed upon SARS-CoV-2 Infection

In order to further validate the role potentially played by the processes identified by GoNet, we analyzed their enrichment for proteins that were found to be differentially expressed upon SARS-CoV-2 infection in human intestinal Caco-2 cells.[27] In this study, Bojkova et al. identified proteins that were significantly differentially expressed 24 h after SARS-CoV-2 infection of Caco-2 cells. From their data, we extracted 309 protein groups that showed significant differential expression (Benjamini–Hochberg FDR-adjusted p-value <0.05). Using a hypergeometric test, we tested whether the proteins annotated with GO terms that were identified as significantly clustered by GoNet in the STRING-extended network were enriched for differential expression (Table 1 and Supplementary Table S8).

Table 1. Gene Ontology Terms for Which the Proteins Are Clustered in the STRING-Extended Network According to GoNet and Are Enriched for Differential Expression upon SARS-CoV-2 Infection

GO identifier	GO name	p-value	FDR-adjusted p-value
GO:0034663	endoplasmic reticulum chaperone complex	5.28×10^{-06}	0.0041
GO:0071407	cellular response to organic cyclic compound	0.00013	0.041
GO:0009628	response to abiotic stimulus	0.00022	0.041
GO:0005924	cell-substrate adherens junction	0.00035	0.041
GO:0005925	focal adhesion	0.00035	0.041
GO:0030055	cell-substrate junction	0.00035	0.041
GO:0030054	cell junction	0.00036	0.041
GO:0005912	adherens junction	0.00050	0.048
GO:0070161	anchoring junction	0.00056	0.048

Table 1 shows that nine GO terms showed such a significant enrichment (FDR-adjusted p-value <0.05). These GO terms are related to "cell junction", "focal adhesion", "response to abiotic stimulus", and the "endoplasmic reticulum chaperone complex", among others. These enrichments show that processes identified by GoNet are potentially significantly affected by SARS-CoV-2, since the viral proteins are tightly interacting with dysregulated proteins involved in such processes.

Protein Sequence Motifs Are Locally Enriched in Certain Regions of the STRING-Augmented and -Extended PPI Networks

In addition to investigating GO overrepresentation in MCL-derived protein clusters from the STRING-augmented and -extended networks, we performed a motif enrichment analysis using the MEME suite (Figure 4, Supplementary File S11 and S12). This analysis highlighted a number of sequence motifs that are surprisingly overrepresented within MCL clusters from both networks (E-value <0.05; Figure 4, Supplementary File S11 and S12). However, the overwhelming majority of these motifs are likely due to high sequence homology between proteins as can be seen from the families to which these proteins belong (i.e., RAB, TIM, TLE, and CSK proteins). Interestingly, a motif involving SRSF proteins, which are involved in regulation of RNA processing, was detected. These

https://dx.doi.org/10.1021/acs.jproteome.0c00422
J. Proteome Res. 2020, 19, 4553–4566

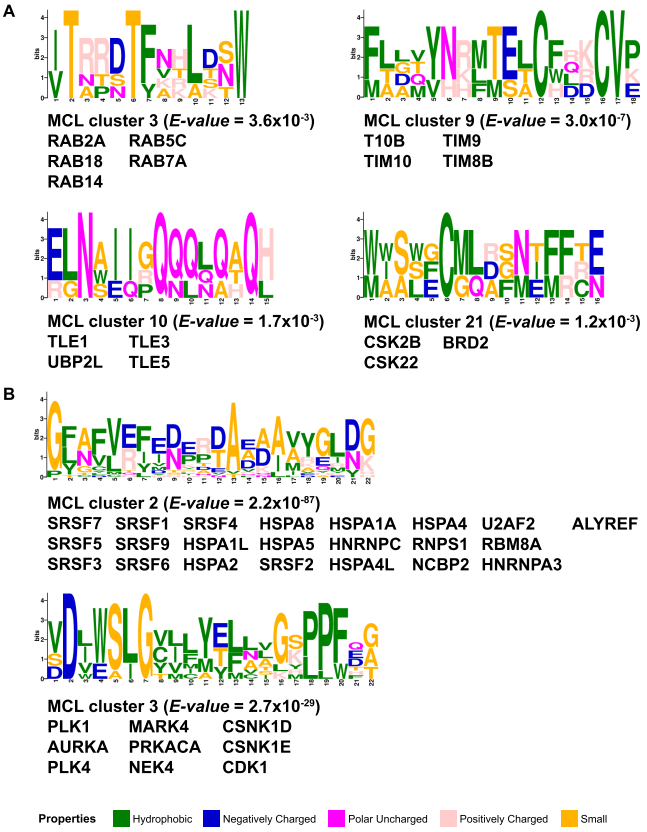


Figure 4. Protein sequence motifs that were enriched in protein clusters detected by MCL in the STRING-augmented PPI network (A) and STRING-extended PPI network (B) (*E*-value <0.05). Proteins containing the motifs are listed below the motif logos. Amino acids are color-coded based on their properties.

motifs may therefore play a role in processes affected by SARS-CoV-2.

LESMoN-Pro Identifies Families of Protein Sequence Motifs That Are Significantly Clustered in the STRING-Augmented and -Extended Networks

Using the LESMoN-Pro algorithm to identify amino acid sequence motifs for which the associated proteins are significantly clustered in the STRING-augmented PPI network, we generated a list of 363 279 motifs (FDR < 0.05). After filtering this list of motifs for which fewer than 75% of the pairs of their associated proteins were homologous (see Experimental Section), 1650 motifs remained (Supplementary Table S9). All of these motifs contained either 3 or 4 degenerate characters. To facilitate downstream analysis of these 1650 motifs and since a large number of motifs are likely to be very similar, we grouped them into 9 families using hierarchical clustering based on the similarity of their lists of associated proteins, and then selected a representative motif for each family (Figure 5 and Supplementary Figure S11). Figure 5A shows that the sets of annotated proteins tended to be centered around highly clustered groups of somewhat homologous proteins, including RAB, NUP, and MARK proteins, but that all included at least one additional protein that share little homology with these families. Supplementary File S13 shows the protein domains in which the motifs were found. Among these proteins, both MIB1 and TBK1, which are involved in the same pathway and implicated in innate antiviral immunity, contained the same representative motif, HXXIXKLb.[28] In the SARS-CoV-2 viral-host PPI network, MIB1 and TBK1 are associated with two different viral proteins, nsp9 and nsp13, respectively, suggesting a potential role for the identified motif in the host response, rather than one directly affecting viral protein binding. Overall, our results identify families of sequence motifs that are significantly clustered in the STRING-augmented PPI network. Such motifs may represent viral protein binding sites and may provide insights into the host processes targeted by SARS-CoV-2 viral proteins and their interactions.

In order to identify protein sequence motifs that are potentially affected by SARS-CoV-2 proteins, we also performed our LESMoN-Pro analysis on the STRING-extended network. This analysis generated a list of 1 121 537 motifs for which the associated proteins were significantly clustered (FDR < 0.005). After filtering out highly homologous motifs, we obtained a list of 8210 motifs (Supplementary Table S10 and File S14) on which we then applied a hierarchical clustering analysis to identify 30 motif families, each defined by a representative motif (Supplementary Figure S12; see Experimental Section). A large fraction of these motifs highly resembles those identified in the STRING-augmented network, thereby confirming their high level of clustering within *H. sapiens* proteins interacting with SARS-CoV-2 proteins and, by the same logic, their potential involvement in infection. For instance, motifs 1 and 2 from the STRING-augmented network (Figure 5) closely resemble motifs 24, 15, 10, and 19 found in the STRING-extended network (Supplementary Figure S12). The same applies for motifs 6 and 9 in the STRING-augmented network (Figure 5), which share similarities with motifs 1, 12, 20, 23, 27, and 29 in the STRING-extended network (Supplementary Figure S12) and motif 5 (STRING-augmented network; Figure 5) which is similar to motifs 8, 11, and 14 (STRING-extended network; Supplementary Figure S12).

Proteins Associated with Motifs Clustered in the STRING-Augmented and -Extended Networks Are Enriched for GO Terms Related to Viral Infection

To further characterize the role potentially played by the motifs in Figure 5 and Supplementary Figure S12, we performed a GO enrichment analysis with Ontologizer on the proteins containing the motifs. The majority of the motifs showed a significant enrichment for at least one GO term (FDR-adjusted p-value (p) < 0.01; Supplementary Figure S13 and Supplementary Files S15 and S16). From these, we note that GO terms related to "transport of virus", "viral life cycle", and "nuclear pore" were enriched among proteins associated with motifs identified in the STRING-augmented network, further suggesting the putative role these motifs are playing in SARS-CoV-2 infection (p < 0.01; Supplementary Figure S13A). On the other hand, proteins associated with motifs identified in the STRING-extended network were enriched for GO terms such as "vesicle", "exocytosis", "cytosolic transport", "membrane", and "virion assembly", again hinting at the role these motifs may play in SARS-CoV-2 infection (p < 0.01; Supplementary Figure S13B).

DISCUSSION

Locally Enriched GO Terms Relevance with Regard to COVID-19

Our initial analysis explored the viral-host PPI network of human proteins interacting with SARS-CoV-2 proteins. This

https://dx.doi.org/10.1021/acs.jproteome.0c00422
J. Proteome Res. 2020, 19, 4553−4566

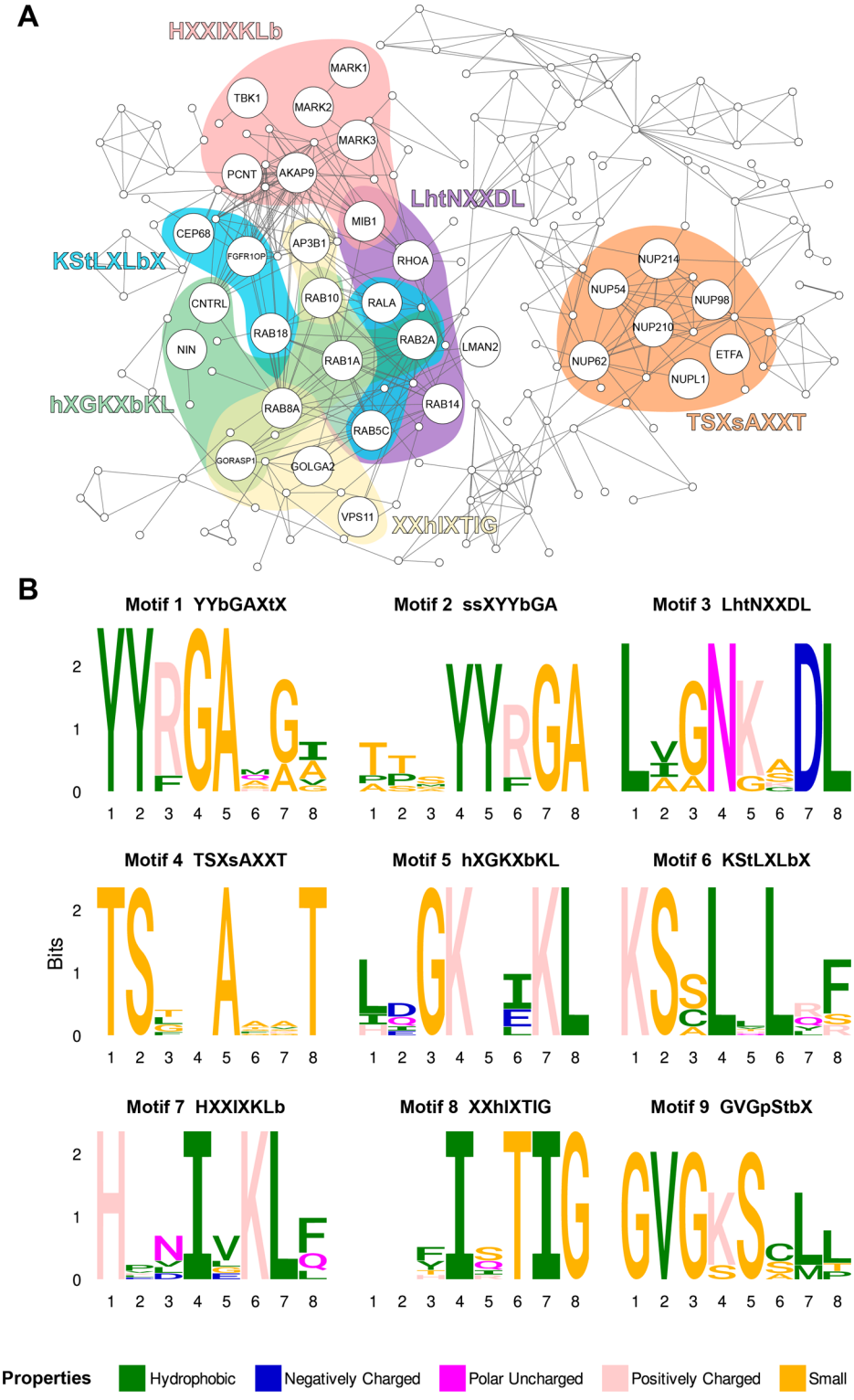


Figure 5. Family representative sequence motifs for which the associated proteins are significantly clustered in the STRING-augmented network. (A) Complete STRING-augmented network where proteins containing significantly clustered motifs are larger and labeled. Selected set of representative motifs are shown on the network coloring the proteins containing them (FDR < 0.05). (B) All family representative motifs detected by LESMoN-Pro are shown as sequence logos built from their actual occurrences in their associated protein sequences. Amino acids are color-coded based on their properties.

analysis highlighted that interactors of the viral nsp proteins were enriched for GO terms such as "MCM complex", "Nucleus", and "Nuclear pore". Given that some nsp proteins have been shown to play an important role in RNA replication and transcription, it is not surprising to see such associations.[29,30] Within this group of nsp proteins, we saw a high enrichment for the "microtubule organizing center" (MTOC) GO term among nsp13 interactors. It has been previously reported that nsp13 is a multifunctional SARS-CoV helicase that unwinds duplex RNA/DNA and, using ATP, is able to

https://dx.doi.org/10.1021/acs.jproteome.0c00422
J. Proteome Res. 2020, 19, 4553−4566

translocate along the nucleotides.[31] It has been an early pharmaceutical target for SARS-CoV inhibition, with the intention of reducing the unwinding ability of the viral helicase.[32] It was proven to be a promising molecular target, as the viral helicase could be inhibited without affecting the activity of the human helicase,[33] and this protein continues to be studied for the novel SARS-CoV-2.[34,35] However, the exact interactions between nsp13 and MTOC have not been characterized to our knowledge. The proteins associated with this GO term may reveal further insight into the replicative mechanism of the virus and assist in the development of improved inhibitors.

We saw a high enrichment for "ubiquitin ligase complex proteins" among the interactors of orf10. This was also confirmed by our GoNet algorithm, which detected ubiquitination-related GO terms as significantly clustered in the network. Proteases have also been an early and promising pharmaceutical target for SARS-CoV-2.[36−38] The entry of the virus has been putatively linked to the ACE2 receptor and priming of the viral S protein by host proteases.[39−41] Interestingly, viral entry can be successfully blocked with the use of a protease inhibitor,[37] making this an exciting intervention option. Thus far, protease-related activity in SARS-CoV-2 has been associated with the viral S protein. Among the novel SARS-CoV-2 proteins, orf10 is one of the least homologous to previous coronavirus proteins[42,43] and therefore has not been as studied to the same extent. However, the association of orf10 to protease-related functions makes this protein an interesting candidate for further investigation. Its implication in the protease pathway can give us further insight into viral protein priming by the host and can ultimately improve current pharmaceutical targets for inhibiting viral entry into the cell.

It is also important to mention that Bojkova et al. reported significant differential expression of spliceosome proteins in their quantitative proteomics study.[27] We also reported using our clustering analysis that RNA processing related GO terms were significantly clustered among the interactors of SARS-CoV-2 proteins (Supplementary Figure S9A). These GO terms represent a few examples that are thought to be useful for therapeutic developments. However, our approach suggests many more GO terms of interest for which the associated proteins may also constitute viable drug targets. Our GO clustering analyses detected clusterings of great significance for annotations such as establishment of RNA localization, cell projection, lyase activity, cell cycle, and more. These fall in line with the current literature that describes such disturbances upon SARS-CoV-2 infection[44−46] and could also represent putative therapeutic targets. This collection of GO terms also helps shed light on the host mechanisms hijacked by SARS-CoV-2.

Locally Enriched Motifs Relevance with Regard to COVID-19

Motifs for which the associated proteins are significantly clustered in the STRING-augmented PPI network may be informative by uncovering potentially important protein sequences involved in both downstream host processes targeted by SARS-CoV-2 and in the direct interactions between viral and host proteins. An example of the former is the pair of proteins MIB1, an E3 ubiquitin-protein ligase, and TBK1, a serine-threonine kinase, both of which share the sequence motif HXXIXKLb and directly interact to regulate innate antiviral immunity by leading to the phosphorylation of interferon regulatory factors, which transcriptionally activate downstream pro-inflammatory and antiviral genes.[28] If SARS-CoV-2 suppresses antiviral immunity in host cells through interactions with these proteins, the analysis of shared sequence motifs may therefore provide insights to discover other proteins that may be involved in these disrupted biological processes. Furthermore, analyzing the clustering of shared sequence motifs in the PPI networks may facilitate the identification of related biological processes that are targeted by, interact with, or affected by SARS-CoV-2 viral proteins. For example, VPS11, a protein involved in vesicle-mediated protein trafficking,[47] AP3B1, a subunit of an adaptor protein complex involved in protein sorting,[48] and GORASP1 and GOLGA2, which are involved in maintenance of[49] and vesicle fusion to[50] the Golgi apparatus, respectively, are significantly clustered in the STRING-augmented PPI network and share the sequence motif XXhIXTIG. This potentially suggests a related role for all four proteins, and possibly the sequence motif itself, in the trafficking of viral components during SARS-CoV-2 infection. Furthermore, one of the motifs reported in Figure 4B is contained by several splicing factors, which are proteins that have been identified by Bojkova et al. to be significantly differentially expressed in Caco-2 cells upon SARS-CoV-2 infection.[27] Finally, while the representative motif chosen for each family provides insights into some of the proteins sharing similar sequence motifs, it is important to note that it does not capture the full variety of the proteins in the network that are annotated by other motifs grouped in the same family. Therefore, further information about the sequence motifs potentially targeted by viral proteins and involved in host processes targeted by SARS-CoV-2 may be derived from further analyses of these additional motif-annotated proteins.

The Influence of Highly Homologous Proteins in Motif Discovery

It is clear that some of the proteins containing the motifs discovered by MEME or LESMoN-Pro share these motifs because they are highly homologous and closely interacting with each other. While these motifs that are part of larger protein sequence domains may be of biological interest to understand the mechanisms of SARS-CoV-2 infection, their occurrences are much less surprising than those in proteins that do not share a high level of homology. For the majority of the motifs reported here, while some proteins hosting these motifs share some level of homology, we observe a good fraction, at least 25% of the proteins, that do not. The facts that these proteins share a high degree of clustering and the presence of a given motif, despite sharing a low level of homology, make them interesting research avenues to provide a better understanding of the SARS-CoV-2 infection. In the future, it could be of interest to mutate such motifs and observe their impact on SARS-CoV-2 infection.

Reliability of PPIs

In this work, we have considered all PPIs to be of equal quality. However, some show a higher level of confidence and some occur at a greater frequency than others. While we have little information about the latter, approaches considering the former to weight PPI networks based on such reliability could be designed. Tools such as SAINT,[51,52] ComPASS,[53] and Decontaminator[54,55] could be used to obtain reliability scores that could be transformed into weights for the interactions in the network. MCL, GoNet, and LESMoN-Pro

https://dx.doi.org/10.1021/acs.jproteome.0c00422
J. Proteome Res. 2020, 19, 4553−4566

could therefore be adapted to consider such weights when performing their clustering analyses to reveal more biologically relevant clusters. While such analyses may be rewarding, it would however remain challenging due to the highly heterogeneous nature of PPIs deposited into STRING.

CONCLUSION

In this manuscript, we report an ensemble of functional annotations, GO terms (biological processes, cellular components, and molecular functions), and Cell Map locations that are likely impacted by SARS-CoV-2 infection in humans. These annotations and their associated proteins can help provide a better understanding of the mechanisms exploited by SARS-CoV-2. We also highlight novel protein sequence motifs that are likely to be closely affected by SARS-CoV-2. Such motifs could represent binding sites of SARS-CoV-2 proteins or play key roles in the processes hijacked by the virus.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.0c00422.

Description of figures, tables, and files (PDF)

Figures S1−S13, Tables S1−S10, Files S1−S16 (ZIP)

AUTHOR INFORMATION

Corresponding Author

Mathieu Lavallée-Adam − *Department of Biochemistry, Microbiology and Immunology, University of Ottawa, Ottawa, Ontario K1H 8M5, Canada;* orcid.org/0000-0003-2124-3872; Phone: 1-613-562-5800; Email: mathieu.lavallee@uottawa.ca

Authors

Rachel Nadeau − *Department of Biochemistry, Microbiology and Immunology, Ottawa Institute of Systems Biology, and Faculty of Medicine, University of Ottawa, Ottawa, Ontario K1H 8M5, Canada*

Soroush Shahryari Fard − *Department of Biochemistry, Microbiology and Immunology, Ottawa Institute of Systems Biology, and Faculty of Medicine, University of Ottawa, Ottawa, Ontario K1H 8M5, Canada*

Amit Scheer − *Department of Biochemistry, Microbiology and Immunology, Ottawa Institute of Systems Biology, and Faculty of Medicine, University of Ottawa, Ottawa, Ontario K1H 8M5, Canada*

Emily Hashimoto-Roth − *Department of Biochemistry, Microbiology and Immunology, Ottawa Institute of Systems Biology, and Faculty of Medicine, University of Ottawa, Ottawa, Ontario K1H 8M5, Canada*

Dallas Nygard − *Department of Biochemistry, Microbiology and Immunology, Ottawa Institute of Systems Biology, and Faculty of Medicine, University of Ottawa, Ottawa, Ontario K1H 8M5, Canada*

Iryna Abramchuk − *Department of Biochemistry, Microbiology and Immunology, Ottawa Institute of Systems Biology, and Faculty of Medicine, University of Ottawa, Ottawa, Ontario K1H 8M5, Canada*

Yun-En Chung − *Department of Biochemistry, Microbiology and Immunology, Ottawa Institute of Systems Biology, and Faculty of Medicine, University of Ottawa, Ottawa, Ontario K1H 8M5, Canada*

Steffany A. L. Bennett − *Department of Biochemistry, Microbiology and Immunology, uOttawa Brain and Mind Research Institute, Ottawa Institute of Systems Biology, uOttawa Brain and Mind Research Institute, Faculty of Medicine, and Department of Chemistry and Biomolecular Sciences, Centre for Catalysis and Research Innovation, University of Ottawa, Ottawa, Ontario K1H 8M5, Canada*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.0c00422

Author Contributions

§R.N., S.S.F., A.S., E.H.R., D.N., I.A., and Y.-E.C. contributed equally to this work.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors acknowledge funding from the following sources: Natural Sciences and Engineering Research Council (NSERC) of Canada Discovery grants to M.L.A. R.N. held a NSERC Canada Graduate Scholarship and an Ontario Graduate Scholarship. S.S.F. held a Canadian Institutes of Health Research Canada Graduate Scholarship. E.H.R. and D.N. were funded by a stipend from the NSERC CREATE in Metabolomics Advanced Training and International Exchange (MATRIX) Program. A.S., I.A., and Y.-E.C. held a NSERC Undergraduate Student Research Award. This research was enabled in part by support provided by Compute Ontario (https://computeontario.ca/) and Compute Canada (www.computecanada.ca) with Resource Allocation to M.L.A.

ABBREVIATIONS

FDR, false discovery rate; GO, gene ontology; MCL, Markov clustering algorithm; MTOC, microtubule organizing center; PPI, protein−protein interaction.

REFERENCES

(1) Wang, C.; Horby, P. W.; Hayden, F. G.; Gao, G. F. A novel coronavirus outbreak of global health concern. *Lancet* **2020**, *395*, 470−473.

(2) Zhu, N.; Zhang, D.; Wang, W.; Li, X.; Yang, B.; Song, J.; Zhao, X.; Huang, B.; Shi, W.; Lu, R.; Niu, P.; Zhan, F.; Ma, X.; Wang, D.; Xu, W.; Wu, G.; Gao, G. F.; Tan, W. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N. Engl. J. Med.* **2020**, *382*, 727−733.

(3) Amanat, F.; Krammer, F. SARS-CoV-2 Vaccines: Status Report. *Immunity* **2020**, *52*, 583−589.

(4) Beigel, J. H.; Tomashek, K. M.; Dodd, L. E.; Mehta, A. K.; Zingman, B. S.; Kalil, A. C.; Hohmann, E.; Chu, H. Y.; Luetkemeyer, A.; Kline, S.; Lopez de Castilla, D.; Finberg, R. W.; Dierberg, K.; Tapson, V.; Hsieh, L.; Patterson, T. F.; Paredes, R.; Sweeney, D. A.; Short, W. R.; Touloumi, G.; Lye, D. C.; Ohmagari, N.; Oh, M.; Ruiz-Palacios, G. M.; Benfield, T.; Fätkenheuer, G.; Kortepeter, M. G.; Atmar, R. L.; Creech, C. B.; Lundgren, J.; Babiker, A. G.; Pett, S.; Neaton, J. D.; Burgess, T. H.; Bonnett, T.; Green, M.; Makowski, M.; Osinusi, A.; Nayak, S.; Lane, H. C. Remdesivir for the Treatment of Covid-19 — Preliminary Report. *N. Engl. J. Med.* **2020**, NEJMoa2007764.

(5) *Summary of Probable SARS Cases with Onset of Illness from 1 November 2002 to 31 July 2003*; World Health Organization, 2003.

(6) *WHO MERS Global Summary and Assessment of Risk, July 2019*; World Health Organization, 2019.

(7) King, A. D.; Przulj, N.; Jurisica, I. Protein complex prediction via cost-based clustering. *Bioinformatics* **2004**, *20*, 3013−3020.
(8) Enright, A. J. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **2002**, *30*, 1575−1584.
(9) Bader, G. D.; Hogue, C. W. V. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinf.* **2003**, *4*, 2.
(10) Gavin, A.-C.; Aloy, P.; Grandi, P.; Krause, R.; Boesche, M.; Marzioch, M.; Rau, C.; Jensen, L. J.; Bastuck, S.; Dümpelfeld, B.; Edelmann, A.; Heurtier, M.-A.; Hoffman, V.; Hoefert, C.; Klein, K.; Hudak, M.; Michon, A.-M.; Schelder, M.; Schirle, M.; Remor, M.; Rudi, T.; Hooper, S.; Bauer, A.; Bouwmeester, T.; Casari, G.; Drewes, G.; Neubauer, G.; Rick, J. M.; Kuster, B.; Bork, P.; Russell, R. B.; Superti-Furga, G. Proteome survey reveals modularity of the yeast cell machinery. *Nature* **2006**, *440*, 631−636.
(11) Sen, T. Z.; Kloczkowski, A.; Jernigan, R. L. Functional clustering of yeast proteins from the protein-protein interaction network. *BMC Bioinf.* **2006**, *7*, 355.
(12) Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M.; Sherlock, G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **2000**, *25*, 25−9.
(13) Ogata, H.; Goto, S.; Sato, K.; Fujibuchi, W.; Bono, H.; Kanehisa, M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **1999**, *27*, 29−34.
(14) Croft, D.; O'Kelly, G.; Wu, G.; Haw, R.; Gillespie, M.; Matthews, L.; Caudy, M.; Garapati, P.; Gopinath, G.; Jassal, B.; Jupe, S.; Kalatskaya, I.; Mahajan, S.; May, B.; Ndegwa, N.; Schmidt, E.; Shamovsky, V.; Yung, C.; Birney, E.; Hermjakob, H.; D'Eustachio, P.; Stein, L. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* **2011**, *39*, D691−7.
(15) Lavallée-Adam, M.; Coulombe, B.; Blanchette, M. Detection of locally over-represented GO terms in protein-protein interaction networks. *J. Comput. Biol.* **2010**, *17*, 443−57.
(16) Lavallée-Adam, M.; Cloutier, P.; Coulombe, B.; Blanchette, M. Functional 5′ UTR motif discovery with LESMoN: Local Enrichment of Sequence Motifs in biological Networks. *Nucleic Acids Res.* **2017**, *45*, 10415−10427.
(17) Szklarczyk, D.; Gable, A. L.; Lyon, D.; Junge, A.; Wyder, S.; Huerta-Cepas, J.; Simonovic, M.; Doncheva, N. T.; Morris, J. H.; Bork, P.; Jensen, L. J.; von Mering, C. STRING v11: protein−protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **2019**, *47*, D607−D613.
(18) Gordon, D. E.; Jang, G. M.; Bouhaddou, M.; Xu, J.; Obernier, K.; White, K. M.; O'Meara, M. J.; Rezelj, V. V.; Guo, J. Z.; Swaney, D. L.; Tummino, T. A.; Huettenhain, R.; Kaake, R. M.; Richards, A. L.; Tutuncuoglu, B.; Foussard, H.; Batra, J.; Haas, K.; Modak, M.; Kim, M.; Haas, P.; Polacco, B. J.; Braberg, H.; Fabius, J. M.; Eckhardt, M.; Soucheray, M.; Bennett, M. J.; Cakir, M.; McGregor, M. J.; Li, Q.; Meyer, B.; Roesch, F.; Vallet, T.; Mac Kain, A.; Miorin, L.; Moreno, E.; Naing, Z. Z. C.; Zhou, Y.; Peng, S.; Shi, Y.; Zhang, Z.; Shen, W.; Kirby, I. T.; Melnyk, J. E.; Chorba, J. S.; Lou, K.; Dai, S. A.; Barrio-Hernandez, I.; Memon, D.; Hernandez-Armenta, C.; Lyu, J.; Mathy, C. J. P.; Perica, T.; Pilla, K. B.; Ganesan, S. J.; Saltzberg, D. J.; Rakesh, R.; Liu, X.; Rosenthal, S. B.; Calviello, L.; Venkataramanan, S.; Liboy-Lugo, J.; Lin, Y.; Huang, X.-P.; Liu, Y.; Wankowicz, S. A.; Bohn, M.; Safari, M.; Ugur, F. S.; Koh, C.; Savar, N. S.; Tran, Q. D.; Shengjuler, D.; Fletcher, S. J.; O'Neal, M. C.; Cai, Y.; Chang, J. C. J.; Broadhurst, D. J.; Klippsten, S.; Sharp, P. P.; Wenzell, N. A.; Kuzuoglu, D.; Wang, H.-Y.; Trenker, R.; Young, J. M.; Cavero, D. A.; Hiatt, J.; Roth, T. L.; Rathore, U.; Subramanian, A.; Noack, J.; Hubert, M.; Stroud, R. M.; Frankel, A. D.; Rosenberg, O. S.; Verba, K. A.; Agard, D. A.; Ott, M.; Emerman, M.; Jura, N.; von Zastrow, M.; Verdin, E.; Ashworth, A.; Schwartz, O.; D'Enfert, C.; Mukherjee, S.; Jacobson, M.; Malik, H. S.; Fujimori, D. G.; Ideker, T.; Craik, C. S.; Floor, S. N.; Fraser, J. S.; Gross, J. D.; Sali, A.; Roth, B. L.; Ruggero, D.; Taunton, J.; Kortemme, T.; Beltrao, P.; Vignuzzi, M.; García-Sastre, A.; Shokat, K. M.; Shoichet, B. K.; Krogan, N. J. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* **2020**, *583*, 459.
(19) Bauer, S.; Grossmann, S.; Vingron, M.; Robinson, P. N. Ontologizer 2.0–a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics* **2008**, *24*, 1650−1.
(20) Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* **1995**, *57*, 289−300.
(21) Go, C. D.; Knight, J. D.; Rajasekharan, A.; Rathod, B.; Hesketh, G. G.; Abe, K. T.; Youn, J.-Y.; Samavarchi-Tehrani, P.; Zhang, H.; Zhu, L. Y. A proximity biotinylation map of a human cell. *bioRxiv*, Oct 7, 2019, 796391. DOI: 10.1101/796391 (accessed Jul 21, 2020).
(22) Bailey, T. L.; Boden, M.; Buske, F. A.; Frith, M.; Grant, C. E.; Clementi, L.; Ren, J.; Li, W. W.; Noble, W. S. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **2009**, *37*, W202−8.
(23) The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **2019**, *47*, D506−D515.
(24) Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T. J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Söding, J.; Thompson, J. D.; Higgins, D. G. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **2011**, *7*, 539.
(25) Wagih, O. ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics* **2017**, *33*, 3645−3647.
(26) El-Gebali, S.; Mistry, J.; Bateman, A.; Eddy, S. R.; Luciani, A.; Potter, S. C.; Qureshi, M.; Richardson, L. J.; Salazar, G. A.; Smart, A.; Sonnhammer, E. L. L.; Hirsh, L.; Paladin, L.; Piovesan, D.; Tosatto, S. C. E.; Finn, R. D. The Pfam protein families database in 2019. *Nucleic Acids Res.* **2019**, *47*, D427−D432.
(27) Bojkova, D.; Klann, K.; Koch, B.; Widera, M.; Krause, D.; Ciesek, S.; Cinatl, J.; Münch, C. Proteomics of SARS-CoV-2-infected host cells reveals therapy targets. *Nature* **2020**, *583*, 469−472.
(28) Li, S.; Wang, L.; Berman, M.; Kong, Y.-Y.; Dorf, M. E. Mapping a Dynamic Innate Immunity Protein Interaction Network Regulating Type I Interferon Production. *Immunity* **2011**, *35*, 426−440.
(29) Keum, Y.-S.; Jeong, Y.-J. Development of chemical inhibitors of the SARS coronavirus: Viral helicase as a potential target. *Biochem. Pharmacol.* **2012**, *84*, 1351−1358.
(30) Romano, M.; Ruggiero, A.; Squeglia, F.; Maga, G.; Berisio, R. A Structural View of SARS-CoV-2 RNA Replication Machinery: RNA Synthesis, Proofreading and Final Capping. *Cells* **2020**, *9*, 1267.
(31) Adedeji, A. O.; Marchand, B.; te Velthuis, A. J. W.; Snijder, E. J.; Weiss, S.; Eoff, R. L.; Singh, K.; Sarafianos, S. G. Mechanism of Nucleic Acid Unwinding by SARS-CoV Helicase. *PLoS One* **2012**, *7*, e36521.
(32) Yu, M.-S.; Lee, J.; Lee, J. M.; Kim, Y.; Chin, Y.-W.; Jee, J.-G.; Keum, Y.-S.; Jeong, Y.-J. Identification of myricetin and scutellarein as novel chemical inhibitors of the SARS coronavirus helicase, nsP13. *Bioorg. Med. Chem. Lett.* **2012**, *22*, 4049−4054.
(33) Subissi, L.; Imbert, I.; Ferron, F.; Collet, A.; Coutard, B.; Decroly, E.; Canard, B. SARS-CoV ORF1b-encoded nonstructural proteins 12−16: Replicative enzymes as antiviral targets. *Antiviral Res.* **2014**, *101*, 122−130.
(34) Shu, T.; Huang, M.; Wu, D.; Ren, Y.; Zhang, X.; Han, Y.; Mu, J.; Wang, R.; Qiu, Y.; Zhang, D.-Y.; Zhou, X. SARS-Coronavirus-2 Nsp13 Possesses NTPase and RNA Helicase Activities That Can Be Inhibited by Bismuth Salts. *Virol. Sin.* **2020**, *35*, 321.
(35) Mirza, M. U.; Froeyen, M. Structural elucidation of SARS-CoV-2 vital proteins: Computational methods reveal potential drug candidates against main protease, Nsp12 polymerase and Nsp13 helicase. *J. Pharm. Anal.* **2020**, *10*, 320.
(36) Wang, J. Fast Identification of Possible Drug Treatment of Coronavirus Disease-19 (COVID-19) through Computational Drug Repurposing Study. *J. Chem. Inf. Model.* **2020**, *60*, 3277.
(37) Hoffmann, M.; Kleine-Weber, H.; Schroeder, S.; Krüger, N.; Herrler, T.; Erichsen, S.; Schiergens, T. S.; Herrler, G.; Wu, N.-H.; Nitsche, A.; Müller, M. A.; Drosten, C.; Pöhlmann, S. SARS-CoV-2

Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **2020**, *181*, 271−280.
(38) Zhang, L.; Lin, D.; Sun, X.; Curth, U.; Drosten, C.; Sauerhering, L.; Becker, S.; Rox, K.; Hilgenfeld, R. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science (Washington, DC, U. S.)* **2020**, eabb3405.
(39) Li, W.; Moore, M. J.; Vasilieva, N.; Sui, J.; Wong, S. K.; Berne, M. A.; Somasundaran, M.; Sullivan, J. L.; Luzuriaga, K.; Greenough, T. C.; Choe, H.; Farzan, M. Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. *Nature* **2003**, *426*, 450−454.
(40) Matsuyama, S.; Nagata, N.; Shirato, K.; Kawase, M.; Takeda, M.; Taguchi, F. Efficient Activation of the Severe Acute Respiratory Syndrome Coronavirus Spike Protein by the Transmembrane Protease TMPRSS2. *J. Virol.* **2010**, *84*, 12658−12664.
(41) Shulla, A.; Heald-Sargent, T.; Subramanya, G.; Zhao, J.; Perlman, S.; Gallagher, T. A Transmembrane Serine Protease Is Linked to the Severe Acute Respiratory Syndrome Coronavirus Receptor and Activates Virus Entry. *J. Virol.* **2011**, *85*, 873−882.
(42) Cagliani, R.; Forni, D.; Clerici, M.; Sironi, M. Coding potential and sequence conservation of SARS-CoV-2 and related animal viruses. *Infect., Genet. Evol.* **2020**, *83*, 104353.
(43) Ahmadpour D, A. P. How the COVID-19 Overcomes the Battle? An Approach to Virus Structure. *Iran. J. Kidney Dis.* **2020**, *14*, 167−172.
(44) da Silva, C. S. B.; Thaler, M.; Tas, A.; Ogando, N. S.; Bredenbeek, P. J.; Ninaber, D. K.; Wang, Y.; Hiemstra, P. S.; Snijder, E. J.; van Hemert, M. J. Suramin inhibits SARS-CoV-2 infection in cell culture by interfering with early steps of the replication cycle. *Antimicrob. Agents Chemother.* **2020**, DOI: 10.1128/AAC.00900-20.
(45) Tang, T.; Bidon, M.; Jaimes, J. A.; Whittaker, G. R.; Daniel, S. Coronavirus membrane fusion mechanism offers a potential target for antiviral development. *Antiviral Res.* **2020**, *178*, 104792.
(46) Fu, Y.; Cheng, Y.; Wu, Y. Understanding SARS-CoV-2-Mediated Inflammatory Responses: From Mechanisms to Potential Therapeutic Tools. *Virol. Sin.* **2020**, *35*, 266.
(47) Brocker, C.; Kuhlee, A.; Gatsogiannis, C.; kleine Balderhaar, H. J.; Honscher, C.; Engelbrecht-Vandre, S.; Ungermann, C.; Raunser, S. Molecular architecture of the multisubunit homotypic fusion and vacuole protein sorting (HOPS) tethering complex. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 1991−1996.
(48) Cowles, C. R.; Odorizzi, G.; Payne, G. S.; Emr, S. D. The AP-3 Adaptor Complex Is Essential for Cargo-Selective Transport to the Yeast Vacuole. *Cell* **1997**, *91*, 109−118.
(49) Hu, F.; Shi, X.; Li, B.; Huang, X.; Morelli, X.; Shi, N. Structural Basis for the Interaction between the Golgi Reassembly-stacking Protein GRASP65 and the Golgi Matrix Protein GM130. *J. Biol. Chem.* **2015**, *290*, 26373−26382.
(50) Puthenveedu, M. A.; Bachert, C.; Puri, S.; Lanni, F.; Linstedt, A. D. GM130 and GRASP65-dependent lateral cisternal fusion allows uniform Golgi-enzyme distribution. *Nat. Cell Biol.* **2006**, *8*, 238−248.
(51) Breitkreutz, A.; Choi, H.; Sharom, J. R.; Boucher, L.; Neduva, V.; Larsen, B.; Lin, Z.-Y.; Breitkreutz, B.-J.; Stark, C.; Liu, G.; Ahn, J.; Dewar-Darch, D.; Reguly, T.; Tang, X.; Almeida, R.; Qin, Z. S.; Pawson, T.; Gingras, A.-C.; Nesvizhskii, A. I.; Tyers, M. A global protein kinase and phosphatase interaction network in yeast. *Science* **2010**, *328*, 1043−6.
(52) Choi, H.; Larsen, B.; Lin, Z.-Y.; Breitkreutz, A.; Mellacheruvu, D.; Fermin, D.; Qin, Z. S.; Tyers, M.; Gingras, A.-C.; Nesvizhskii, A. I. SAINT: probabilistic scoring of affinity purification-mass spectrometry data. *Nat. Methods* **2011**, *8*, 70−3.
(53) Sowa, M. E.; Bennett, E. J.; Gygi, S. P.; Harper, J. W. Defining the Human Deubiquitinating Enzyme Interaction Landscape. *Cell* **2009**, *138*, 389−403.
(54) Lavallée-Adam, M.; Rousseau, J.; Domecq, C.; Bouchard, A.; Forget, D.; Faubert, D.; Blanchette, M.; Coulombe, B. Discovery of cell compartment specific protein-protein interactions using affinity purification combined with tandem mass spectrometry. *J. Proteome Res.* **2013**, *12*, 272−281.
(55) Lavallée-Adam, M.; Cloutier, P.; Coulombe, B.; Blanchette, M. Modeling contaminants in AP-MS/MS experiments. *J. Proteome Res.* **2011**, *10*, 886−95.
(56) Supek, F.; Bošnjak, M.; Škunca, N.; Šmuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* **2011**, *6*, e21800.
(57) Kuznetsova, I.; Lugmayr, A.; Siira, S. J.; Rackham, O.; Filipovska, A. CirGO: an alternative circular way of visualising gene ontology terms. *BMC Bioinf.* **2019**, *20*, 84.

https://dx.doi.org/10.1021/acs.jproteome.0c00422
J. Proteome Res. 2020, 19, 4553−4566