Reconstructing SARS-CoV-2 response signaling and regulatory networks

Jun Ding*[1], Jose Lugo-Martinez*[1], Ye Yuan*[2], Darrell N. Kotton[3,4], and Ziv Bar-Joseph†[1,2]

[1]Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, 15213, USA
[2]Machine Learning Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, 15213, USA
[3]Center for Regenerative Medicine of Boston University and Boston Medical Center, Boston, MA 02118, USA
[4]The Pulmonary Center and Department of Medicine, Boston University School of Medicine, Boston, MA 02118, USA

June 1, 2020

Abstract

Several molecular datasets have been recently compiled to characterize the activity of SARS-CoV-2 within human cells. Here we extend computational methods to integrate several different types of sequence, functional and interaction data to reconstruct networks and pathways activated by the virus in host cells. We identify the key proteins in these networks and further intersect them with genes differentially expressed at conditions that are known to impact viral activity. Several of the top ranked genes do not directly interact with virus proteins though some were shown to impact other coronaviruses highlighting the importance of large scale data integration for understanding virus and host activity.

1 Introduction

To fully model the impact of the novel coronavirus, SARS-CoV-2, which causes the COVID-19 pandemic, requires the integration of several different types of molecular and cellular data. SARS-CoV-2 is known to primarily impact cells via two viral entry factors, ACE2 and TMPRSS2 [1]. However, much less is currently known about the virus activity once it enters lung cells. Similar to other viruses, once it enters the cell response to SARS-CoV-2 leads to the activation and repression of several pro and anti inflammatory pathways and networks [2]. Predictive mechanistic models about the activity of these pathways, the key proteins that participate in them and the regulatory networks that these pathways activate is the first step towards identifying useful drug targets.

Recently, several studies provided information about various aspects of the molecular activity of SARS-CoV-2. These include studies focused on inferring virus-host interactions [3], studies focused on the viral sequence and mutations [4], studies profiling expression changes following infection [5]

*These authors have contributed equally to this work.
†Correspondence: Ziv Bar-Joseph, zivbj@cs.cmu.edu

and studies identifying underlying health conditions that lead to an increased likelihood of infection and death [6, 7]. While these studies are very informative, each only provides a specific view-point regarding virus activity. By integrating all of this data in a single model we can enhance each of the specific data types and reconstruct the networks utilized by the virus and host following infection.

Several unsupervised methods have been developed for inferring molecular networks from interaction and expression data [8, 9]. However, these usually cannot link the expression changes observed following infection to the viral proteins that initiate the responses. To enable such analysis that directly links sources (viral proteins and their human interactors) with targets (genes that are activated / repressed following infection) we used the SDREM method [10, 11], which reconstructs regulatory and signaling networks and identifies key proteins that mediate infection signals. SDREM combines Input-Output Hidden Markov models (IOHMM) with graph directionality assignments to link sources and targets. It then performs combinatorial network analysis to identify key nodes that can block the signals between sources and targets and ranks them to identify potential targets for treatments. Here we extended SDREM so that it can utilize static expression data and take into account the expression of viral genes as well.

The extended method was applied to three gene expression datasets profiling responses to SARS-CoV and SARS-CoV-2. As we show, the networks reconstructed using all three datasets were in very good agreement and identified a number of very relevant genes as potential targets. Intersecting the list of top-scoring SDREM genes with genes identified as differentially expressed (DE) in several of the underlying conditions further narrows the list of candidate targets. Many of the top predictions are not directly interacting with viral proteins and so cannot be identified without integrating several different data types.

2 Methods

2.1 Datasets

We used both condition specific and general interaction data for learning dynamic viral infection models. Some of the condition specific data was from SARS-CoV-2 experiments while for others we combined SARS-CoV and SARS-CoV-2 data. We also collected lung expression data for conditions that were reported to impact SARS-CoV-2 mortality and infection rates. Below we provide information about all data used in this study.

2.1.1 Viral host interactions

We used the SARS-CoV-2 and human protein-protein interactome reported by Gordon *et al.* [3] which identified 332 protein interactions between 26 viral proteins and 332 human proteins using affinity purification-mass spectrometry analysis. While this data is virus specific, prior studies for other viruses indicated that a single screen is unlikely to fully cover the entire set of virus-host interactions [12]. In order to increase the CoV-human interactome coverage, we also downloaded protein sequences for other coronaviruses (e.g., SARS-CoV and MERS-CoV) from the NCBI GenBank database (18 March 2020) and identified protein homologs to the SARS-CoV-2 viral proteins. Next, we compiled additional human proteins associated with each viral protein from VirHostNet 2.0 [13], BioGRID [14] and IntAct [15] as well as published literature. The resulting set of CoV–human protein-protein interaction network included 459 protein interactions between 34 CoV proteins and 443 human proteins. Supplementary Table 1 provides the full list of interactions we used.

2.1.2 RNAi knockdown data

We searched the literature for a list of RNAi screen experiments which test the impact of gene knockdown/knockout on coronavirus load. In particular, we collected RNAi screening data for 4 different coronaviruses: IBV-CoV, MERS-CoV, MHV-CoV and SARS-CoV. The combined set of RNAi screen hits is comprised of 272 human proteins in cells infected with a coronavirus, 6 of which were present in our initial SARS-CoV-2 and human network. Table 1 summarizes the hits used in this study while Supplementary Table 2 provides the full list of screen hits used in this study.

Coronavirus	Screen hits	References
IBV-CoV	94	[16, 17]
MERS-CoV	20	[18, 19]
MHV-CoV	26	[20, 21, 22]
SARS-CoV	135	[23, 24, 25]

Table 1: **Summary of literature-derived coronaviruses RNAi screen hits**. For each screen, we list associated coronavirus, number of significant RNAi hits and corresponding references. We use 10 different RNAi screen studies for avian infectious bronchitis virus (IBV-CoV), Middle East respiratory syndrome (MERS-CoV), murine hepatitis virus (MHV-CoV) and severe acute respiratory syndrome (SARS-CoV) coronaviruses.

2.1.3 General Protein-protein and protein-DNA interactions

Protein-DNA interactions were obtained from our previous work [26], which contains 59578 protein-DNA interactions for 399 Transcription Factors (TFs). Protein-protein interactions (PPIs) were obtained from the HIPPIE database [27], which contains more than 270000 annotated PPIs and for each provides a confidence score which was further used in our network analysis (see below).

2.1.4 Time series and static transcriptomics infection data

We used two different types of transcriptomics data to learn SDREM SARS models. The first was a time series RNA-seq data of SARS-CoV infection of Calu-3 cells, a human lung epithelial adenocarcinoma cell line previously employed in preclinical in vitro modeling [28]. Samples were collected at 0, 3, 7, 12, 24, 30, 36, 48, 54, 60, and 72 hours post infection. We also download static human SARS-CoV-2 infection RNA-seq data from BioProject Accession # PRJNA616446. Two bronchoalveolar lavage samples were collected, and the gene expression (for both virus and host) was profiled using RNA-seq. Of the ∼12.4M reads For the first RNA-seq sample (S9), 7.55% were mapped to viral genome and 22.27% reads were mapped to exons, the majority of the remaining reads are likely bacteria genomes. There were a total of ∼11.7M reads for the second RNA-seq sample (S10). Of these 0.7% reads were mapped to the viral genome and 27.55% reads were mapped to human exons.

2.1.5 Underlying condition lung expression data

We collected and analyzed lung expression data from several reported underlying conditions that impact SARS-CoV-2 infection. Table 2 provides information on these conditions, the impact they were reported to have and the source and type of expression data we analyzed. Most of the data we used was from bulk microarray expression studies of lung tissues. These included studies focused on lung cancer [29], hypertension [30], Diabetes [31], Chronic Obstructive Pulmonary Disease

(COPD) [32], and smoking [33]. We also used single cell RNA-seq lung data for gender expression analysis and for inferring aging related genes from [34]. For each dataset, the differentially expressed genes are extracted using the R package 'limma' for microarray data [35] and by using a ranksum test for single cell data.

Condition	Accession number	Data type	References
Smoker	GSE10072	Microarray	[33]
Hypertension	GSE24988	Microarray	[30]
Diabetes	GSE15900	Microarray	[31]
COPD	GSE38974	Microarray	[32]
Cancer	GSE2514	Microarray	[29]
Age	GSE136831	scRNA-seq	[34]
Sex	GSE136831	scRNA-seq	[34]

Table 2: **Summary of literature-derived condition lung expression data**. For each condition, we list the accession number, data type and corresponding references.

2.2 Reconstructing dynamic signaling and regulatory networks using SDREM

For the analysis and modeling of SARS-CoV-2 infection in lung cells, we extend the Signaling Dynamic Regulatory Events Miner (SDREM) [11] method. SDREM integrates time-series gene expression data with static PPIs and protein-DNA interaction to reconstruct response regulatory networks and signaling pathways. SDREM iterates between two methods. The first, DREM [36] uses an input-output hidden Markov model (IOHMM) to reconstruct dynamic regulatory networks by identifying bifurcation events where a set of co-expressed genes diverges. DREM annotates these splits and paths (co-expressed genes) with TFs that regulate genes in the outgoing upward and/or downward paths. The second part of SDREM uses a network orientation method [10], which orients the undirected protein interaction edges such that the targets can be explained by relatively short, high-confidence pathways that originated at the inputs with provided PPIs, source proteins, and target TFs. Generally, SDREM searches for high scoring paths that start at the virus proteins, continues with the host proteins they interact with and ends with TFs and their targets. By iterating between identifying TFs based on the expression of their targets and connecting identified TFs to source (virus) proteins, SDREM can identify a set of high scoring pathways and regulators. These are then analyzed using graph based scoring methods as we discuss below to identify key proteins mediating viral signals.

2.3 Using human SARS-CoV-2 data with SDREM

We modified SDREM in two ways to improve its performance on the SARS-CoV-2 data. First, much of the infection data we used is static. To enable the use of SDREM on such data we first combined the infection RNA-Seq data with control RNA-seq for lung alveolar cells from [37]. This enabled us to identify DE genes between infected and healthy cells. To remove the potential batch effect, we have performed cross-sample normalization between the bulk RNA-seq expression of normal alveolar cells and SARS-CoV-2 infected cells. We next inserted a psuedo-time point 0 (with no change) and used the log fold change of infected vs. normal cells as the second time point value. See Figure 1 for the resulting SDREM model for this data. Another extension we applied is the use of the viral expression data to remove sources and their connected host proteins. Specifically, we first identified the non-expressed viral genes in those datasets. Then, we remove all the host

source proteins that interact with the viral proteins that correspond to those non-expression viral genes.

2.4 Identifying key genes in SDREM reconstructed networks

To rank top genes identified by SDREM, we used the strategy described in [10] to estimate *in silico* effects of removing a protein from the signaling network component of an SDREM model. The method computes how the connectivity to the TFs is affected when a node (gene) or a combination of nodes, is removed. Intuitively, this score captures the impact of the removal on the path weights that remain for linking TFs to sources (eqn.1).

$$score_w(A) = \frac{\sum_{t \in T} \frac{\sum_{p \in P(t)} w(p) I(A \in N(p)))}{\sum_{p \in P(t)} w(p)}}{|T|} \tag{1}$$

Where A is the deleted node, T is the set of all targets, $P(t)$ is the set of paths to the target t to be considered, $w(p)$ is the path weight, $I(*)$ is an indicator function that has the value 1 if the condition * is satisfied, $N(p)$ is the set of nodes on the path p.

Although single gene inference might be very informative, higher-order knockdowns (of two or more genes) may prove to be more robust because they can target several pathways simultaneously. Experimentally testing all possible gene combinations would be prohibitive but in-silico analysis is much faster given the relatively small number of genes in the resulting SDREM network. Scores for pair removals are computed in a similar way to individual scores, by finding double knockdown has a stronger effect than expected based on the score for individual gene of the pair, which corresponds to lower value of $P_{AB}^{observed}$ (see eqn.2 for details).

$$P_{AB}^{observed} = score_w(A, B) = \frac{\sum_{t \in T} \frac{\sum_{p \in P(t)} w(p) I(A \notin N(P)) I(B \notin N(p))}{\sum_{p \in P(t)} w(p)}}{|T|} \tag{2}$$

The above eqn.2 denotes the average fraction of path weight that remains after removing paths that contain node A and B.

Several rankings can be derived based on the score computed above. These differ in the paths used for the scoring (top ranked or all), weather target connectivity is evaluated separately for every source or for all sources combined and using weighted VS. unweighted versions of the SDREM network. See Supplementary Table S3 (Meta) for details on what was used in the analysis.

2.5 Analyzing underlying condition data

We examined several recent studies to determine underlying condition that impact SARS-CoV-2 mortality and infection rates [6, 7]. Based on these we selected seven different conditions for which we were able to obtain lung expression data (bulk and / or single cell). For cancer, hypertension, Diabetes, COPD, and smoking, we used microarray lung expression data as shown in Table 2. For each dataset we compared the case samples with normal samples to obtain a set of DE genes and their (adjusted) p-value using 'limma' [35]. For the 'age' and 'gender' factors, we used scRNA-Seq data. For each sample, we averaged the expression values for each gene in all cells of the sample to get expression data form similar to bulk data. As for 'age', we first ordered samples based on their age, and selected samples younger than 60 as control, and older ones as case samples. For gender, we only focused on samples younger than 60, and used female as control to perform DE

analysis for the male samples. We next used the ranksum statistical test to compute DE p-value for each gene. Finally, we used the assigned p-values and expression fold change signs to compute quantile value between 0 and 1 for each gene in each condition and divided genes in top quantile for each condition to 'over-expressed' with label of 1, 'repressed' with label of -1, and non-differentially expressed gene with label of 0 (See Supplementary Table S5 for details).

2.6 Intersecting SDREM and underlying condition genes

We intersected the list of top ranked genes from the two SDREM reconstructed networks with genes identified as significantly up or down regulated in the underlying conditions data. The intersection was computed for several combinations of conditions and p-values and summarized in Figure 2A-B. For each entry in that table we computed the p-value for the intersection represented in that entry using the hypergeometric distribution.

2.6.1 Potential treatments for top genes

In a similar fashion to Gordon *et al.* [3], we searched public resources (e.g., IUPHAR/BPS and ChEMBL25 [38]), as well as published literature in order to identify existing drugs and reagents that directly modulate the candidate genes derived from our network reconstruction and condition-specific analyses.

3 Results

We integrated several condition specific and general molecular datasets to reconstruct infection pathways for SARS-CoV-2 in lung cells. While SARS-CoV-2 is known to primarily impact cells via two viral entry factors, ACE2 and TMPRSS2, much less is currently known about the virus activity within lung cells. To model the pathways it activates and represses and their potential impact on viral load we start with host protein known to interact with virus proteins [3], and attempt to link them through signaling pathways to the observed dynamic regulatory network following viral infection. By analyzing the top ranked nodes in the reconstructed networks we can identify the proteins that are likely to play key roles in either increasing or decreasing viral loads. We further intersected these top ranked proteins with up and down regulated genes in a large cohort of lung expression data from conditions that are known to impact SARS-CoV-2 mortality rates.

3.1 Reconstructed signaling and regulatory network

We analyzed the signaling and regulatory networks of the SARS-CoV and SARS-CoV-2 RNA-seq samples using SDREM. We first compiled the initial coronavirus-host protein-protein interaction network by combining the human proteins that interact with SARS-CoV-2 proteins with the two key viral entry proteins, ACE2 and TMPRSS2 (Methods). For the SARS-Cov RNA-seq dataset, the learned SDREM network contained 155 proteins, which include 123 source proteins (host proteins that interact with viral proteins) and 32 non-source proteins of which 21 are TFs. The network reconstructed using SARS-CoV-2 RNA-seq dataset S9, included 158 proteins with 117 source proteins and 41 non-source proteins comprising 30 TFs. For the SARS-CoV-2 RNA-seq dataset S10, the SDREM network contained 160 proteins (121 source and 39 non-source including 28 TFs). See Figure 1 and Supplementary Table S3 for the detailed SDREM predicted protein networks.

Intersection of the networks reconstructed for the two SARS-CoV-2 samples (S9 and S10) identified 122 proteins (over 78.7%) as shared between the two. Importantly, the intersection

included 26 of 41 (63.4%) non-source proteins for S9. Since unlike source proteins non-source are selected from over 15000 potential proteins, the overlap is significant (p-value=0) and indicates that the networks agree on key interactions. 10 of the 30 (33.3%) predicted TFs were also identified for both datasets. Of the 122 proteins, 112 (91.8%) are also shared with the SARS-CoV data result including 16 of 26 non-source proteins (61.5%). 7 TFs are shared across all networks (Supplementary Table S4).

3.2 Identifying top ranked SDREM genes

We next ranked all proteins identified for the different SARS-CoV-2 expression datasets (S9 and S10, Methods) and computed intersections for the top 100 ranked proteins (plus a few target TFs that are not scored). We found that 79 proteins are in the top 100 for both S9 and S10 including 17 non-source proteins. 65 proteins that are ranked in the top 100 for all 3 datasets, 10 of which are non-source proteins. We next analyzed the functions of predicted proteins. Since source proteins were used as input, the analysis focused on the newly identified non-sources and TFs. We found that some of the most enriched GO terms for the shared non-source proteins include positive regulation of viral transcription (FDR=7.75e-3), regulation of signaling (9.87e-5), and immune response-regulating cell surface receptor signaling pathway (2.43e-3). See Supplementary Table S5 for complete GO term list.

SDREM can also be used to rank protein pairs. 17577 protein pairs were ranked for the SARS-CoV-2 S9 dataset, and 18914 protein pairs were ranked for the SARS-CoV-2 S10 dataset. 15752 (∼90%) protein pairs are shared between the two networks. 227 pairs that include 154 protein (138 source and 16 signaling) proteins are in the top 1000 list for both sets. The 16 non-source proteins in the intersection are enriched with several relevant GO terms including: negative regulation of RNA biosynthetic process (FDR=1.8e-08), negative regulation of transcription of RNA polymerase II (2.68e-08), and Viral process (8.02e-06). See Supplementary Table S5 for complete GO term list.

3.3 Activity of top ranked genes in underlying health conditions

To further narrow down the list of potential host genes that impact viral loads and activity we next looked at the activity of these genes in a set underlying health conditions that were determined to impact SARS-CoV-2 infection and mortality rates. Table 2 presents the 7 conditions we looked at. For each of these we identified one or more lung gene expression dataset and ranked genes by the DE score (up or down, Methods) for each condition. We next intersected the different sets of top DE genes from each condition with top 100 ranked genes and the selected target TFs for the S9 and S10 networks (Methods). Results are presented in Figure 2. We identified several parameters for which the intersection is significant meaning that many of the top SDREM genes are also DE in several of the conditions. We focused on an number of parameters. For overlap with top S9-S10 individual list, using a quantile threshold of 0.3 and sum threshold of 3 results in 19 genes with (p value = 7.38e-6, hypergeometric distribution), as shown in Figure 2C. Some of the genes in the intersection are TFs that are expressed in many tissue (e.g. JUN and MYC) and are also known to have important roles in proliferation in lung epithelial cells. PLK1 is another interesting gene on the list which has been previously linked to proliferation and migration in non small cell lung cancers, a tumor type arising from lung epithelial cells [39, 40].

More generally the list of genes is enriched for GO terms that include 'viral process' (p value = 3.92e-6), 'symbiotic' (p value = 5.31e-7), and a series of 'cell cycle' (minimal p value = 1.48e-7) relevant functions. We also computed overlap for each model (S9, S10) separately, see Supplementary Figures S1 and S2 for details. For overlap between SARS-associated SDREM gene list and

condition gene list (Figure S3), the same quantile and sum thresholds result in 26 genes (p value = 1.03e-7), with enriched GO terms of that include 'viral process' (p value = 3.05e-8), 'symbiotic'(p value = 5.80e-9) and 'cell cycle' (p value = 1.54e-8).

We also analyzed the top pairs for each model. As mentioned above, 227 pairs were in the overlap of the top 1000 for S9 and S10 containing a total of 154 genes. We analyzed the overlap between these 154 genes and the condition specific genes. Using a quantile threshold of 0.3, sum threshold of 3, we obtained an overlap of 39 overlap genes (p value = 4.35e-11), as shown in Figure 3 for details. Two pairs, (BRCA1, EP300) and (BRCA1, JUN) were present in both, the pair overlap list and the top condition gene list.

3.4 Intersection with RNAi knockdown studies

We compared the set of shared proteins between the two SARS-CoV-2 samples (122 proteins, S9 and S10) against RNAi screen hits for multiple coronaviruses (Methods). We identified 5 proteins (p value = 2.57e-2) which have been previously shown to affect coronavirus load in RNAi screen experiments. It is worth noting that 2 genes on this list (PLK1 and UBE2I) are also identified as condition related genes and thus appear in all relevant datasets we analyzed (network analysis, condition related and RNAi). The smaller list of shared proteins between all models (SARS-CoV and SARS-CoV-2, 112 proteins) includes 3 proteins (p value = 1.96e-1) which have been annotated to alter coronavirus replication across different RNAi experiments. Of these 3 proteins, RAB7A is very interesting since it is a lysosomal-endosomal protein in many cell types and in alveolar epithelial type 2 cells has an important role in disease pathogenesis and likely is part of both the endosomal and lamlellar body-multivesicular body organelles whose normal function is required for proper surfactant packaging and secretion in the lung [41]. Table 3 summarizes the identified proteins supported by RNAi experiments.

Gene name	SARS-CoV-2 interactor?	RNAi supported effect
Overlap between both SARS-CoV-2 samples		
BCKDK	Y	increased SARS-CoV replication
PLK1	N	significantly decreased MERS-CoV replication
RAB7A	Y	reduced MHV-CoV replication
UBE2I	N	decrease IBV-CoV replication
VPS11	Y	severely reduced MHV-CoV infection
Overlap between SARS-CoV and SARS-CoV-2 samples		
BCKDK	Y	increased SARS-CoV replication
RAB7A	Y	reduced MHV-CoV replication
VPS11	Y	severely reduced MHV-CoV infection

Table 3: **Summary of proteins supported by RNAi screen experiments identified by SDREM**. For each list, we denote whether it is a known interactor of a SARS-CoV-2 protein and a brief description of the experimental impact on coronavirus load. Top table enumerates each protein previously reported to affect coronavirus load in RNAi screen experiments for the shared proteins between the two SARS-CoV-2 samples (p value = 2.57e-2), whereas bottom table lists the proteins in RNAi screen hits for shared proteins between all SARS models (p value = 1.96e-1). Proteins also listed in the top 100 of each overlap are shown in boldface, p values: 2.27e-2 (top) and 2.21e-1 (bottom).

3.5 Potential treatments for predicted genes

We looked for potential treatments for the 39 proteins identified at the intersection of top ranked SDREM and underlying condition genes. Table 4 lists 10 human proteins that we identified as potential drug targets using curated databases of bioactive molecules such as ChEMBL and IUPHAR/BPS Guide to Pharmacology. Among these potential drug targets, 60% (i.e, 6 out of 10) are not characterized as SARS-CoV-2 interactors and 40% (i.e, 4 out of 10) are also not known as CoV interactors (i.e. non-source in our networks). It is worth highlighting that Gordon *et al.* [3] tested antiviral activity of 47 compounds targeting known SARS-CoV-2 interactors. Of these drug targets tested for inhibition of viral infection, only BRD2 is also identified in our list, unfortunately, the antiviral activity results were inconclusive. Supplementary Table 6 provides the full list of chemical associations to human proteins identified as potential drug targets.

Gene name	SARS-CoV-2 interactor?	CoV interactor?
BRCA1	N	N
BRD2	Y	Y
EP300	N	N
JUN	N	Y[a]
MYC	N	N
NEU1	Y	Y
PLK1	N	N
SLC27A2	Y	Y
TBK1	Y	Y
UBE2I	N	Y

Table 4: **Summary of identified proteins associated with chemical compounds in ChEMBL25 or IUPHAR/BPS Guide to Pharmacology**. For each human protein, we show whether or not it is a known SARS-CoV-2 or CoV interactor, respectively. Proteins further associated with drugs and reagents are shown in boldface. [a] denotes that protein is not a direct target of CoV proteins but involved in crucial pathways of CoV infection [42]

.

4 Discussion

By integrating data from several relevant molecular resources we were able to identify a subset of genes that are (1) connected to viral proteins in signaling pathways (2) impact downstream expression response to the infection and (3) identified to be DE in underlying conditions. We used computational methods that combine probabilistic graphical models with combinatorial network analysis to rank top genes and pairs of genes and to intersect these with underlying condition genes.

Our methods identified a list of 19 genes in the overlap of all relevant datasets when looking at top single node rankings and 39 genes when looking at pair rankings. Functional analysis of the these genes indicated that many are related to host response to viral infections and to replications, the two key types of pathways expected to be activated following infection. While some of the top ranked proteins are well known, several are novel predictions that have not been previously studied since they do not directly interact with SARS-CoV-2 proteins. 10 of the proteins in our intersection sets have known potential treatments. Of these, 6 are not characterized as SARS-CoV-2 interactors and 4 are not known as to interact with any CoV virus. Thus, these proteins have not been previously highlighted as potential molecular factors of SARS-CoV-2 treatment. This

list includes PLK1 and UBE2I which, while not directly interacting with a CoV were identified as RNAi screen hits for CoV highlighting them as likely candidates for further experimental evaluation. Moreover, the two gene pairs (BRCA1, EP300) and (BRCA1, JUN) identified in the intersection of pairs and DE genes are also potentially interesting drug targets as each protein has several high-affinity compounds associations, thus, suggesting a drug mixture to treat multiple targets.

While our analysis utilizes all current available data, it is important to note that the bulk SARS-CoV-2 gene expression data we used was from Bronchoalveolar lavage and so may be mostly comprised of immune rather than airway cells. As more expression datasets are released we intend to expand our model by applying it to the increasing amounts and types of gene expression data profiled following infection.

The methods we presented is general and can be applied to integrate several additional molecular studies focused on SARS-CoV-2. Since each of the data types (interactions, expression, knockdown etc.) provides a unique and complementary viewpoint about the virus activity, a model that can integrate the data to reconstruct a mechanistic model is required in order to obtain insights that will help identify potential treatments.

References

[1] M. Hoffmann, H. Kleine-Weber, S. Schroeder, N. Krüger, T. Herrler, S. Erichsen, T. S. Schiergens, G. Herrler, N.-H. Wu, A. Nitsche, *et al.*, "SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor," *Cell*, 2020.

[2] Y. Fu, Y. Cheng, and Y. Wu, "Understanding SARS-CoV-2-mediated inflammatory responses: from mechanisms to potential therapeutic tools," *Virol Sin*, pp. 1–6, 2020.

[3] D. E. Gordon, G. M. Jang, M. Bouhaddou, J. Xu, *et al.*, "A SARS-CoV-2 protein interaction map reveals targets for drug repurposing," *Nature*, pp. 1–44, 2020.

[4] T. Phan, "Genetic diversity and evolution of SARS-CoV-2," *Infec Genet Evol*, vol. 81, p. 104260, 2020.

[5] Y. Xiong, Y. Liu, L. Cao, D. Wang, M. Guo, A. Jiang, D. Guo, W. Hu, J. Yang, Z. Tang, *et al.*, "Transcriptomic characteristics of bronchoalveolar lavage fluid and peripheral blood mononuclear cells in COVID-19 patients," *Emerg Microbes Infect*, vol. 9, no. 1, pp. 761–770, 2020.

[6] X. Zhao, B. Zhang, P. Li, C. Ma, J. Gu, P. Hou, Z. Guo, H. Wu, and Y. Bai, "Incidence, clinical characteristics and prognostic factor of patients with COVID-19: a systematic review and meta-analysis," *medRxiv*, 2020.

[7] F. Zhou, T. Yu, R. Du, G. Fan, Y. Liu, Z. Liu, J. Xiang, Y. Wang, B. Song, X. Gu, *et al.*, "Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study," *Lancet*, 2020.

[8] A. I. Huynh-Thu, V. A., L. Wehenkel, and P. Geurts, "Inferring regulatory networks from expression data using tree-based methods," *PLoS One*, vol. 5, no. 9, 2010.

[9] S. Aibar, C. B. González-Blas, T. Moerman, H. Imrichova, G. Hulselmans, F. Rambow, J.-C. Marine, P. Geurts, J. Aerts, J. van den Oord, *et al.*, "SCENIC: single-cell regulatory network inference and clustering," *Nat Methods*, vol. 14, no. 11, pp. 1083–1086, 2017.

[10] A. Gitter and Z. Bar-Joseph, "Identifying proteins controlling key disease signaling pathways," *Bioinformatics*, vol. 29, no. 13, pp. i227–i236, 2013.

[11] A. Gitter, M. Carmi, N. Barkai, and Z. Bar-Joseph, "Linking the signaling cascades and dynamic regulatory networks controlling stress responses," *Genome Res*, vol. 23, no. 2, pp. 365–376, 2013.

[12] F. D. Bushman, N. Malani, J. Fernandes, I. D'Orso, G. Cagney, T. L. Diamond, H. Zhou, D. J. Hazuda, A. S. Espeseth, R. König, S. Bandyopadhyay, T. Ideker, S. P. Goff, N. J. Krogan, A. D. Frankel, J. A. T. Young, and S. K. Chanda, "Host cell factors in HIV replication: Meta-analysis of genome-wide studies," *PLoS Pathog*, vol. 5, pp. 1–12, 2009.

[13] T. Guirimand, S. Delmotte, and V. Navratil, "VirHostNet 2.0: surfing on the web of virus/host molecular interactions data," *Nucleic Acids Res*, vol. 43, no. D1, pp. D583–D587, 2014.

[14] R. Oughtred, C. Stark, B.-J. Breitkreutz, J. Rust, *et al.*, "The BioGRID interaction database: 2019 update," *Nucleic Acids Res*, vol. 47, no. D1, pp. D529–D541, 2018.

[15] S. Orchard, M. Ammari, B. Aranda, L. Breuza, *et al.*, "The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases," *Nucleic Acids Res*, vol. 42, no. D1, pp. D358——D363, 2014.

[16] Y. W. Tan, W. Hong, and D. X. Liu, "Binding of the 5'-untranslated region of coronavirus RNA to zinc finger CCHC-type and RNA-binding motif 1 enhances viral replication and transcription," *Nucleic Acids Res*, vol. 40, no. 11, pp. 5065–5077, 2012.

[17] H. H. Wong, P. Kumar, F. P. L. Tay, D. Moreau, D. X. Liu, and F. Bard, "Genome-wide screen reveals valosin-containing protein requirement for coronavirus exit from endosomes," *J Virol*, vol. 89, no. 21, pp. 11116–11128, 2015.

[18] J. Kindrachuk, B. Ork, B. J. Hart, S. Mazur, M. R. Holbrook, M. B. Frieman, D. Traynor, R. F. Johnson, J. Dyall, J. H. Kuhn, G. G. Olinger, L. E. Hensley, and P. B. Jahrling, "Antiviral potential of ERK/MAPK and PI3K/AKT/mTOR signaling modulation for middle east respiratory syndrome coronavirus infection as identified by temporal kinome analysis," *Antimicrob Agents Chemother*, vol. 59, no. 2, pp. 1088–1099, 2015.

[19] S. Dirmeier, C. Dächert, M. van Hemert, A. Tas, N. S. Ogando, F. van Kuppeveld, R. Bartenschlager, L. Kaderali, M. Binder, and N. Beerenwinkel, "Host factor prioritization for pan-viral genetic perturbation screens using random intercept models and network propagation," *PLoS Comput Biol*, vol. 16, no. 2, pp. 1–19, 2020.

[20] K. S. Choi, A. Mizutani, and M. M. C. Lai, "SYNCRIP, a member of the heterogeneous nuclear ribonucleoprotein family, is involved in mouse hepatitis virus RNA synthesis," *J Virol*, vol. 78, no. 23, pp. 13153–13162, 2004.

[21] M. W. Vogels, B. W. M. van Balkom, D. V. Kaloyanova, J. J. Batenburg, A. J. Heck, J. B. Helms, P. J. M. Rottier, and C. A. M. de Haan, "Identification of host factors involved in coronavirus replication by quantitative proteomics analysis," *Proteomics*, vol. 11, no. 1, pp. 64–80, 2011.

[22] C. Burkard, M. H. Verheije, O. Wicht, S. I. van Kasteren, F. J. van Kuppeveld, B. L. Haagmans, L. Pelkmans, P. J. M. Rottier, B. J. Bosch, and C. A. M. de Haan, "Coronavirus cell

entry occurs through the endo-/lysosomal pathway in a proteolysis-dependent manner," *PLoS Pathog*, vol. 10, pp. 1–17, 11 2014.

[23] N. Yang, P. Ma, J. Lang, Y. Zhang, J. Deng, X. Ju, G. Zhang, and C. Jiang, "Phosphatidylinositol 4-kinase iiiβ is required for severe acute respiratory syndrome coronavirus spike-mediated cell entry," *J Biol Chem*, vol. 287, no. 11, pp. 8457–8467, 2012.

[24] A. H. de Wilde, K. F. Wannee, F. E. M. Scholte, J. J. Goeman, P. ten Dijke, E. J. Snijder, M. Kikkert, and M. J. van Hemert, "A kinome-wide small interfering rna screen identifies proviral and antiviral host factors in severe acute respiratory syndrome coronavirus replication, including double-stranded RNA-activated protein kinase and early secretory pathway proteins," *J Virol*, vol. 89, no. 16, pp. 8318–8333, 2015.

[25] Y. Ma-Lauer, J. Carbajo-Lozoya, M. Y. Hein, M. A. Müller, W. Deng, J. Lei, B. Meyer, Y. Kusov, B. von Brunn, D. R. Bairad, S. Hünten, C. Drosten, H. Hermeking, H. Leonhardt, M. Mann, R. Hilgenfeld, and A. von Brunn, "p53 down-regulates SARS coronavirus replication and is targeted by the SARS-unique domain and PL^{pro} via E3 ubiquitin ligase RCHY1," *Proc Natl Acad Sci*, vol. 113, no. 35, pp. E5192–E5201, 2016.

[26] J. Ding, B. J. Aronow, N. Kaminski, J. Kitzmiller, J. A. Whitsett, and Z. Bar-Joseph, "Reconstructing differentiation networks and their regulation from time series single-cell expression data," *Genome Res*, vol. 28, no. 3, pp. 383–395, 2018.

[27] G. Alanis-Lobato, M. A. Andrade-Navarro, and M. H. Schaefer, "HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks," *Nucleic Acids Res*, vol. 45, no. D1, pp. D408–D414, 2016.

[28] A. C. Sims, S. C. Tilton, V. D. Menachery, L. E. Gralinski, A. Schäfer, M. M. Matzke, B.-J. M. Webb-Robertson, J. Chang, M. L. Luna, C. E. Long, *et al.*, "Release of severe acute respiratory syndrome coronavirus nuclear import block enhances host transcription in human lung cells," *J Virol*, vol. 87, no. 7, pp. 3885–3902, 2013.

[29] R. S. Stearman, L. Dwyer-Nield, L. Zerbe, S. A. Blaine, Z. Chan, P. A. Bunn Jr, G. L. Johnson, F. R. Hirsch, D. T. Merrick, W. A. Franklin, *et al.*, "Analysis of orthologous gene expression between human pulmonary adenocarcinoma and a carcinogen-induced murine model," *Am J Pathol*, vol. 167, no. 6, pp. 1763–1775, 2005.

[30] M. Mura, M. Anraku, Z. Yun, K. McRae, M. Liu, T. K. Waddell, L. G. Singer, J. T. Granton, S. Keshavjee, and M. de Perrot, "Gene expression profiling in the lungs of patients with pulmonary hypertension associated with pulmonary fibrosis," *Chest*, vol. 141, no. 3, pp. 661–673, 2012.

[31] E. van Lunteren, M. Moyer, and S. Spiegler, "Alterations in lung gene expression in streptozotocin-induced diabetic rats," *BMC Endocr Disord*, vol. 14, no. 1, p. 5, 2014.

[32] M. E. Ezzie, M. Crawford, J.-H. Cho, R. Orellana, S. Zhang, R. Gelinas, K. Batte, L. Yu, G. Nuovo, D. Galas, *et al.*, "Gene expression networks in COPD: microRNA and mRNA regulation," *Thorax*, vol. 67, no. 2, pp. 122–131, 2012.

[33] M. T. Landi, T. Dracheva, M. Rotunno, J. D. Figueroa, H. Liu, A. Dasgupta, F. E. Mann, J. Fukuoka, M. Hames, A. W. Bergen, *et al.*, "Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival," *PLoS One*, vol. 3, no. 2, 2008.

[34] T. S. Adams, J. C. Schupp, S. Poli, E. A. Ayaub, N. Neumark, F. Ahangari, S. G. Chu, B. A. Raby, G. DeIuliis, M. Januszyk, *et al.*, "Single cell RNA-seq reveals ectopic and aberrant lung resident cell populations in idiopathic pulmonary fibrosis," *BioRxiv*, p. 759902, 2019.

[35] M. E. Ritchie, B. Phipson, D. I. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, "limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic Acids Res*, vol. 43, no. 7, pp. e47–e47, 2015.

[36] J. Ding, J. S. Hagood, N. Ambalavanan, N. Kaminski, and Z. Bar-Joseph, "iDREM: Interactive visualization of dynamic regulatory networks," *PLoS Comput Biol*, vol. 14, no. 3, p. e1006019, 2018.

[37] J. Ding, F. Ahangari, C. R. Espinoza, D. Chhabra, T. Nicola, X. Yan, C. V. Lal, J. S. Hagood, N. Kaminski, Z. Bar-Joseph, *et al.*, "Integrating multiomics longitudinal data to reconstruct networks underlying lung development," *Am J Physiol Lung Cell Mol Physiol*, vol. 317, no. 5, pp. L556–L568, 2019.

[38] D. Mendez, A. Gaulton, A. P. Bento, J. Chambers, M. De Veij, E. F/'elix, M. P. Magari/ nos, J. F. Mosquera, P. Mutowo, M. Nowotka, M. Gordillo-Marañón, F. Hunter, L. Junco, G. Mugumbate, M. Rodriguez-Lopez, F. Atkinson, N. Bosc, C. J. Radoux, A. Segura-Cabrera, A. Hersey, and A. R. Leach, "ChEMBL: towards direct deposition of bioassay data," *Nucleic Acids Res*, vol. 47, no. D1, pp. D930–D940, 2018.

[39] W. Yan, H. Yu, W. Li, *et al.*, "Plk1 promotes the migration of human lung adenocarcinoma epithelial cells via STAT3 signaling," *Oncol Lett*, vol. 16, no. 5, pp. 6801–6807, 2018.

[40] J. A. Stratmann and M. Sebastian, "Polo-like kinase 1 inhibition in NSCLC: mechanism of action and emerging predictive biomarkers," *Lung Cancer (Auckl)*, vol. 10, pp. 67–80, 2019.

[41] D. R. Hawkins A., Guttentag S.H. *et al.*, "A non-BRICHOS SFTPC mutant (SP-CI73T) linked to interstitial lung disease promotes a late block in macroautophagy disrupting cellular proteostasis and mitophagy," *Am J Physiol Lung Cell Mol Physiol*, vol. 308, no. 1, pp. L33–L47, 2015.

[42] Y. Zhou, Y. Hou, J. Shen, Y. Huang, W. Martin, and F. Cheng, "Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2," *Cell Discov*, vol. 6, no. 1, p. 14, 2020.

Figures

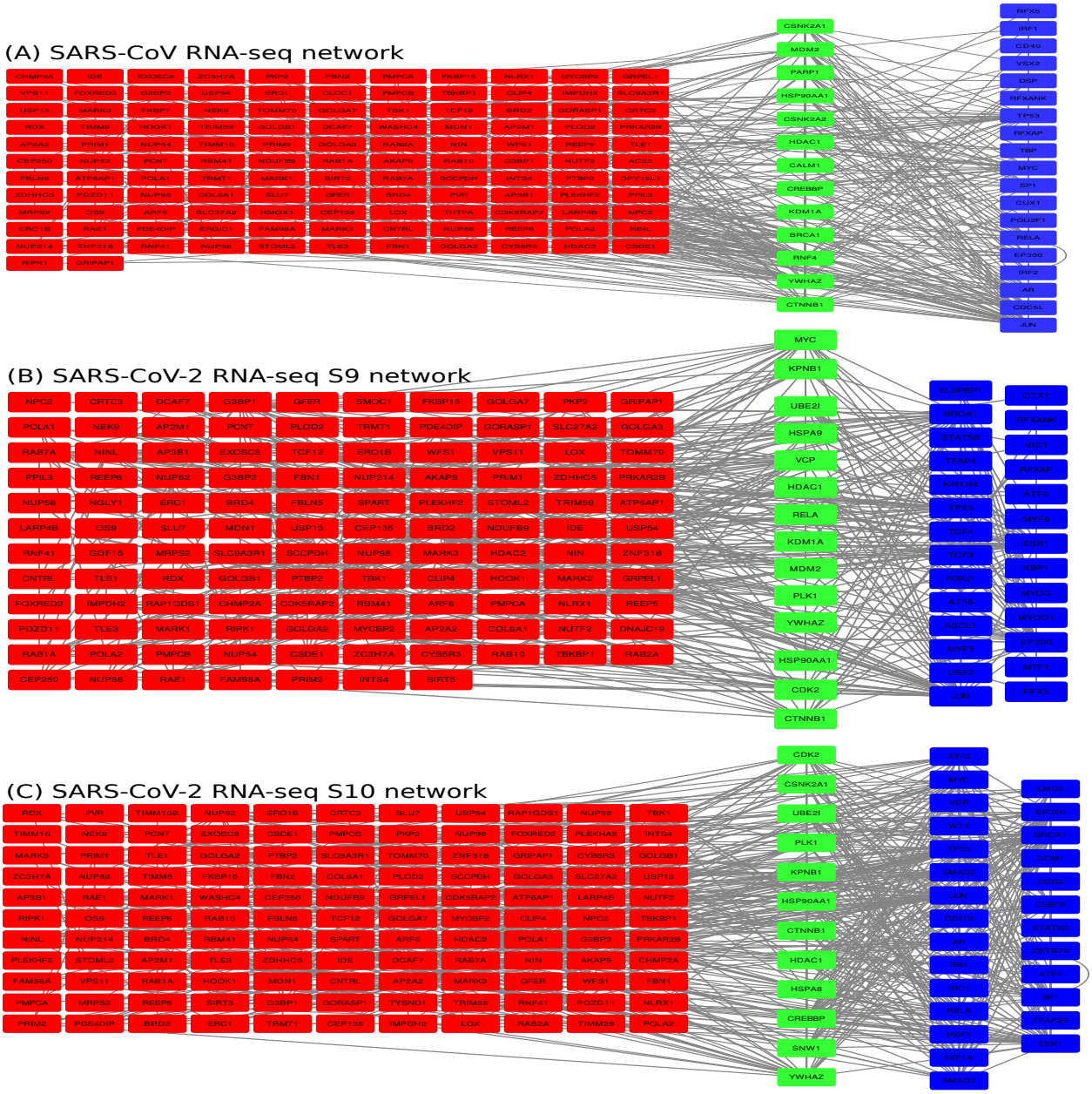

Figure 1: SDREM predicted networks for 3 RNA-seq datasets. Red nodes denote sources, green nodes are signaling proteins and blue nodes are TFs associated with regulating the DREM portion of the model. Overlap between the networks is discussed in Results.

A

Quantile threshold \ Sum score threshold	0	1	2	3	4	5	6
0.01	0.531712	2.30E-05	0.131169	1	1	1	1
0.05	0.531712	4.25E-06	0.001675	1	1	1	1
0.1	0.531712	0.003358	0.000269	0.018208	0.143289	0.018587	1
0.15	0.531712	0.017335	3.86E-05	0.005602	0.3578	0.066461	1
0.2	0.531712	0.03547	0.003413	0.000153	0.053498	1	1
0.25	0.531712	0.01824	0.000187	0.000213	0.012688	0.250807	1
0.3	0.531712	0.041083	2.78E-05	7.38E-06	0.036782	0.397452	0.109194

B

Quantile threshold \ Sum score threshold	0	1	2	3	4	5	6
0.01	79	18	1	0	0	0	0
0.05	79	48	8	0	0	0	0
0.1	79	51	16	4	1	1	0
0.15	79	55	24	7	1	1	0
0.2	79	59	23	12	3	0	0
0.25	79	66	32	14	5	1	0
0.3	79	66	40	19	5	1	1

C

gene	hypertension	COPD	diabetes	smoke	cancer	sex	age	sum
ap2a2	-1	-1	-1	-1	0	-1	-1	-6
Jun	-1	1	-1	-1	-1	-1	0	-4
csde1	1	-1	0	-1	0	-1	-1	-3
myc	-1	1	0	0	-1	-1	-1	-3
reep5	0	-1	0	-1	-1	-1	1	-3
ube2i	0	-1	-1	-1	-1	0	1	-3
brd2	-1	1	-1	-1	-1	-1	1	-3
tle1	0	-1	0	-1	-1	-1	1	-3
tbk1	-1	1	1	0	0	1	1	3
ep300	1	1	1	-1	-1	1	1	3
pcnt	1	-1	1	0	0	1	1	3
ngly1	1	-1	0	1	0	1	1	3
dcaf7	1	0	0	1	1	-1	1	3
nup214	1	0	0	1	-1	1	1	3
plk1	1	1	-1	1	1	-1	1	3
prim2	1	1	0	1	1	-1	0	3
fam98a	0	1	0	0	1	1	1	4
brca1	1	0	0	1	1	1	0	4
slc9a3r1	1	1	-1	0	1	1	1	4

D

GO biological process complete	raw P value	FDR
mitotic cell cycle	1.48E-07	2.37E-03
symbiotic process	5.31E-07	4.24E-03
mitotic cell cycle process	1.29E-06	6.85E-03
beta-catenin-TCF complex assembly	3.12E-06	1.25E-02
viral process	3.92E-06	1.25E-02
regulation of cell cycle	6.67E-06	1.78E-02
cellular macromolecule metabolic process	9.93E-06	1.98E-02
negative regulation of gene expression	9.90E-06	2.26E-02
cell cycle	1.39E-05	2.47E-02
protein modification by small protein conjugation or removal	1.65E-05	2.63E-02
cell cycle process	2.47E-05	3.28E-02
regulation of mitotic cell cycle	2.38E-05	3.45E-02
signal transduction involved in cell cycle checkpoint	4.48E-05	4.47E-02
signal transduction involved in DNA integrity checkpoint	4.31E-05	4.59E-02
response to estrogen	5.22E-05	4.90E-02
signal transduction involved in DNA damage checkpoint	4.31E-05	4.92E-02
negative regulation of macromolecule metabolic process	4.20E-05	5.16E-02

Figure 2: Overlap analysis between condition and SDREM individual gene list. (A) P values for overlap genes with different quantile and sum score threshold combination in condition gene set. (B) number of overlap genes. (C) The selected 19 genes with quantile threshold of 0.3 and sum threshold of 3. (D) Significant GO terms for (C).

A

Sum score threshold

	0	1	2	3	4	5	6
0.01	0.52	9.58E-07	0.030918	1	1	1	1
0.05	0.52	8.62E-07	1.95E-06	0.075732	1	1	1
0.1	0.52	1.63E-05	7.84E-09	3.38E-05	0.25908	0.035724	1
0.15	0.52	0.000193	2.22E-10	1.45E-06	0.056613	0.124836	1
0.2	0.52	0.000441	2.41E-08	2.62E-10	1.37E-06	0.038139	1
0.25	0.52	0.006288	5.04E-09	2.74E-10	4.56E-06	2.69E-05	0.084146
0.3	0.52	0.004073	6.27E-09	4.35E-11	8.35E-05	0.00349	0.02153

B

Sum score threshold

	0	1	2	3	4	5	6
0.01	154	30	2	0	0	0	0
0.05	154	79	17	2	0	0	0
0.1	154	103	35	10	1	1	0
0.15	154	114	51	17	3	1	0
0.2	154	125	56	28	11	2	0
0.25	154	122	68	33	13	6	1
0.3	154	131	78	39	13	5	2

C

gene	hypertension	COPD	diabetes	smoke	cancer	sex	age	sum
ap2a2	-1	-1	-1	-1	0	-1	-1	-6
fbln5	-1	-1	-1	-1	-1	0	0	-5
wfs1	0	-1	-1	-1	-1	0	-1	-5
prkar2b	0	-1	-1	-1	-1	-1	1	-4
jun	-1	1	-1	-1	-1	-1	0	-4
csde1	1	-1	0	-1	0	-1	-1	-3
gpx1	0	-1	0	0	0	-1	-1	-3
myc	-1	1	0	0	-1	-1	-1	-3
cyb5r3	-1	-1	-1	-1	-1	1	1	-3
golgb1	1	-1	-1	-1	-1	-1	1	-3
reep5	0	-1	0	-1	-1	-1	1	-3
reep6	0	0	-1	0	0	-1	-1	-3
tbkbp1	0	-1	0	-1	-1	-1	1	-3
usp54	1	-1	-1	0	0	-1	-1	-3
ube2i	0	-1	-1	-1	-1	0	1	-3
ergic1	-1	-1	-1	0	0	-1	1	-3
brd2	-1	1	-1	-1	-1	-1	1	-3
tle1	0	-1	0	-1	-1	-1	1	-3
tbk1	-1	1	1	0	0	1	1	3
golga3	0	1	0	-1	1	1	1	3
ep300	1	1	1	-1	-1	1	1	3
rab1a	0	1	1	-1	1	0	1	3
gcc2	1	1	1	-1	0	0	1	3
pcnt	1	-1	1	0	0	1	1	3
ngly1	1	-1	0	1	0	1	1	3
dcaf7	1	0	0	1	1	-1	1	3
nup214	1	0	0	1	-1	1	1	3
plk1	1	1	-1	1	1	-1	1	3
trim59	1	1	0	0	0	0	1	3
neu1	-1	1	1	1	1	-1	1	3
prim2	1	1	0	1	1	-1	0	3
fam98a	0	1	0	0	1	1	1	4
nptx1	1	-1	0	1	1	1	1	4
brca1	1	0	0	1	1	1	0	4
fbn2	1	1	1	1	0	1	-1	4
nup58	-1	1	1	1	1	1	0	4
slc9a3r1	1	1	-1	0	1	1	1	4
slc27a2	1	0	1	0	1	1	1	5
ero1b	1	1	1	1	0	1	1	6

D

GO biological process complete	raw P value	▲ FDR
symbiotic process	6.63E-08	1.05E-03
viral process	1.96E-06	1.56E-02
positive regulation of protein modification by small protein conjugation or removal	8.41E-06	4.46E-02

Figure 3: Overlap analysis between condition and SDREM gene pair list. (A) P values for overlap genes . (B) number of overlap genes. (C) The selected 39 genes with quantile threshold of 0.3 and sum threshold of 3. (D) Significant GO terms for (C).