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

October 10, 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.

Software and interactive visualization: https://github.com/phoenixding/sdremsc

---

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

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, 4], studies focused on the viral sequence and mutations [5], studies profiling expression changes following infection [6] and studies identifying underlying health conditions that lead to an increased likelihood of infection and death [7, 8]. While these studies are very informative, each only provides a specific viewpoint 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 [9, 10]. 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 extended SDREM method [11, 12], 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 phosphorylation, time series single cell expression data and information on the expression of viral genes.

The extended method was applied to two gene expression datasets profiling responses to SARS-CoV-2 in bulk and single cells. As we show, the networks reconstructed using these datasets were in very good agreement and have also agreed with independent proteomics studies. The networks identified several 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.

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 proteins known to interact with virus proteins [3, 4], and attempt to link them through signaling pathways to the observed dynamic regulatory network following viral infection. Next, our method ranks all proteins based on their participation in pathways that link sources (proteins directly interacting with the virus) and targets (expressed genes) and selects top ranking pathways. These are used to reconstruct networks which contain proteins and interactions 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.

Reconstructed signaling and regulatory network from bulk transcriptomics

We used our SDREM method to reconstruct signaling and regulatory networks activated following SARS-CoV-2 infection. We first combined virus-host and host-host protein interaction data with bulk time series expression data of lung epithelial cells infected by SARS-CoV-2 (Methods). The reconstructed network is presented in Supplementary Figure S1. The signaling part contained 252 proteins, of which 203 are source proteins (host proteins that interact with viral proteins) and 49 are proteins that do not directly interact with viral proteins. These 49 proteins included 34 TFs

that directly regulate the expression downstream target genes and 15 internal (signaling) proteins which serve as the intermediate between the source proteins and the TFs.

We first examined the 203 source proteins selected by the method. These represent only 17.7% of the 1148 human proteins that were experimentally identified as interacting with virus proteins (and that served as input to the method). The most enriched GO term (using ToppGene [13]) for the 203 selected source proteins is 'viral process' (FDR=2.178e-14). In contrast, no GO category related to viruses is found to be significant for the full set of 1148 source proteins. This indicates that by integrating several diverse datasets SDREM is able to zoom in on the most relevant source proteins from the two studies.

The 15 internal proteins are also enriched for relevant functions including 'cellular response to steroid hormone stimulus (FDR=1.153e-9)' which is also significantly associated with the 34 target TFs (FDR=1.666e-12). The combined set of 49 non-source proteins are also significantly enriched with sex hormone response related functions. The role of sex hormone in COVID-19 was recently reported [14].

We further validated the proteins identified by SDREM by comparing them to 543 proteins that were determined to be phosphorylated following infection with SARS-CoV-2 [15]. We observed a significant overlap between SDREM selected proteins and the list of phosphorylated proteins (19 out of 252 proteins are phosphorylated: hyper-geometric test p-value=6.35e-5).

Integrating phosphorylation data to reconstruct networks

While the phosphorylation data can serve as a validation, we can also use it as an input to increase the prior for including a protein in the reconstructed network (Methods). Given the significant intersection between the networks reconstructed without such data and the phosphorylation data we next reconstructed networks that use this data in addition to the protein-protein, protein-DNA interaction and time series expression data.

Networks learned with this data included a total of 261 proteins with 204 source proteins 17 internal proteins and 40 TFs (Figure 1). We again observe 'viral process' as the most enriched category for the source proteins (Supplementary Table S1). (FDR=1.173e-10). The 17 internal proteins are significantly enriched with transcription relevant GO terms such as "negative regulation of transcription, DNA-templated (FDR=2.772e-7)" and sex hormone stimulus relevant functions such

as "intracellular steroid hormone receptor signaling pathway (FDR=2.772e-7)". The 40 target TFs are enriched with "positive regulation of transcription by RNA polymerase II (FDR=1.066E-39)" and "response to hormone (FDR=2.738E-12)", which are consistent with the function of internal proteins.

In addition to identifying proteins that play a major role in key pathways, SDREM can also be used to identify pairs of proteins that, together, control a significant number of pathways (i.e., pairs that are expected to have the largest impact in a double KO experiment, Methods). For the learned network SDREM also identified 28712 protein pairs. These protein pairs are composed of 243 distinct proteins (203 sources, 17 internal, 22 TFs), which slightly differ from the proteins identified based on their individual rankings (243/261=93.% single knock-out proteins are also found in double knock-out analysis). See Supplementary Table S1 for the complete list of protein pairs.

Reconstructed signaling and regulatory network from single-cell transcriptomics

To further narrow down the list of key signaling and regulatory factors involved in SARS-CoV-2 response we next used SDREM to analyze time series scRNA-Seq SARS-CoV-2 infection data (Methods). The signaling part reconstructed for this data contained 244 proteins, of which 172 are source proteins (host proteins that interact with viral proteins) and 72 are proteins that do not directly interact with viral proteins. These 72 proteins included 49 TFs that directly regulate the expression downstream target genes and 23 internal (signaling) proteins which serve as the intermediate between the source proteins and the TFs. We found that the majority of the 244 proteins (159/65.2%, p-value=2.26e-12) are shared between the single-cell and bulk reconstructed networks (135, source proteins 14 TFs, and 10 internal proteins). Figure 2 presents the conserved signaling network at the intersection of both models. The most enriched GO term (using ToppGene) for the 135 selected source proteins is 'viral process' (FDR=1.01e-11). The enriched GO terms for the 24 non-source proteins include 'regulation of transcription'(FDR=3.23e-19), and 'response to cytokine' (FDR=6.250E-15). Please refer to Supplementary Table S2 for the detailed GO terms.

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 1 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 source and internal genes found by both the bulk and single cell SDREM network and with selected target TFs (Methods). The intersection results are presented in Figure 3 (See Supplementary Figures S2 and S3 for bulk SDREM result with phosphorylation and Supplementary Figures S4 and S5 for single-cell SDREM result with phosphorylation). 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. For overlap with top individual list, using a quantile threshold of 0.25 and sum threshold of 2 results in 33 genes with (p value = 2.63e-6, hypergeometric distribution), as shown in Figure 3C. Some of the genes in the intersection are TFs that are expressed in many tissue (e.g., JUN and FOS) and are also known to have important roles in lung. CAV 1 is another interesting gene which is essential in the acute lung injury pathogenesis [16]. More generally the list of genes is enriched for GO terms that include 'cellular response to chemical stimulus' (p value = 1.80e-11), 'viral process' (p value = 2.37e-10), transcription and metabolism relevant functions.

We also focused on the list of top 1000 pairs identified as described above. There were 91 unique genes in the overlap top pair list from bulk SDREM and single cell SDREM analysis, and we analyzed the intersection between them and the condition specific genes. Using a quantile threshold of 0.25, sum threshold of 3, we obtained an overlap of 21 genes (p value = 1.57e-7), as shown in Figure 4. This list of genes is enriched for GO terms that include 'response to external stimulus' (p value = 5.08e-8), 'nitrogen compound metabolic process' (p value = 9.34e-8) and several other immune, metabolism and transcription relevant functions.

Intersection with RNAi or CRISPR knockdown studies

We next compared the 261 proteins identified by SDREM from bulk RNA-seq expression with RNAi and CRISPR screens for multiple coronaviruses (Methods). We identified 16 proteins (p value = 6.32e-4) which have been previously shown to affect coronavirus load in RNAi screen experiments (Table 2). It is worth noting that 4 genes on this list (ATM, CAV1, SMAD3 and UBE2I) are also identified as condition related genes (p value = 2.82e-2) and thus appear in all relevant datasets we analyzed (network analysis, condition related and RNAi). The smaller list of top 1000 protein pairs identified by SDREM (168 proteins) includes 10 proteins (p value = 7.70e-3) which have been annotated to alter coronavirus replication across different RNAi experiments. Of these 10 proteins, RAB7A is a potentially interesting target. RAB7A is a lysosomal-endosomal protein that is found in alveolar epithelial type 2 (AT2) cells. RAB7A has an important role in disease pathogenesis and 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 [17]. In addition to the 4 genes previously identified as condition-associated by the single ranking method the pairs method also identified EPHA2 as a condition related gene (p value = 2.14e-2). Table 2 summarizes the identified proteins supported by RNAi or CRISPR experiments.

In a similar manner, we also analyzed the set of 244 proteins identified by SDREM from single-cell RNA-seq expression data against the same list of RNAi and CRISPR screens across coronaviruses (Methods). Table 2 also lists the 14 proteins identified as previously shown to affect coronavirus load in RNAi or CRISPR screen experiments (p value = 2.50e-3). Interestingly, 11 proteins (~78%) are identified by both reconstructed networks (bulk and sc, Table 2). Again, 4 out of the 14 genes on this list (ATM, CAV1, CSNK2A1 and SMAD3) are also identified as condition related genes (p value = 7.82e-2). Among these genes, all except CSNK2A1 are also identified in the bulk network. Finally, the list of top 1000 protein pairs identified by SDREM (217 proteins) includes the same 14 proteins (p value = 8.28e-4) which have been annotated to affect coronavirus replication across different RNAi and CRISPR experiments.

7

Potential treatments for predicted genes

We looked for potential treatments for 90 total proteins (48 from bulk analysis and 67 from single-cell analysis) identified at the intersection of top ranked SDREM and underlying condition genes. Table 3 lists 32 human proteins that we identified as potential drug targets using curated databases of bioactive molecules such as ChEMBL, Pharos and ZINC. Among these potential drug targets, 56% (i.e., 18 out of 32) are not characterized as SARS-CoV-2 interactors (i.e., non-source in our networks). Additionally, among the 32 human proteins, 8 proteins (25%) are identified separately from bulk and single-cell RNA-seq data. 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 and CSNK2A2 are also identified in our list. They reported viral inhibition for CSNK2A2 with compound Silmitasertib (CX-4945), whereas the antiviral activity results for BRD2/4 across 6 different compounds were inconclusive. Supplementary Table S3 provides the full list of chemical associations to human proteins identified as potential drug targets.

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 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. As shown in Table 3, 32 of the proteins in our intersection sets have known potential treatments, including 9 proteins associated with approved drugs (EGFR, ERBB2, ESRRA, IRAK1, JAK2, NR3C1, PLAU, RELA and RXRA). Of these, 5

are not characterized as SARS-CoV-2 interactors. Thus, these proteins have not been previously highlighted as potential molecular factors of SARS-CoV-2 treatment. Additionally, the full list of potential drug targets includes CSNK2A1 and UBE2I which, while not directly interacting with a CoV, are identified as RNAi or CRISPR screen hits for CoV highlighting them as likely candidates for further experimental evaluation. Moreover, any gene pair from Table 3, where one protein in the pair has an associated approved drug and the other protein has several high-affinity compounds associations, for example (ERBB2, EP300) or (ERBB2, MYC), are also potentially interesting therapeutic targets, thus, suggesting a drug mixture to treat multiple targets.

While our analysis utilizes a time series SARS-CoV-2 infection data, it only contains 2 time points following infection which may miss a number of critical events. 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.

Materials and Methods

Datasets

We used both condition specific and general interaction data for learning dynamic viral infection models. 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.

Viral host interactions and phosphorylation data

We used the SARS-CoV-2 and human protein-protein interactome reported by Gordon *et al.* [3] and Stukalov *et al.* [4] which identified 1396 protein interactions between 31 viral proteins and 1148 human proteins using affinity purification-mass spectrometry analysis. While this data is virus specific, we note that prior studies for other viruses indicated that a single screen is unlikely to fully cover the entire set of virus-host interactions [18]. Supplementary Table S4 provides the full list of interactions we used. We also used a protein phosphorylation data dataset in which phosphorylation levels were profiled at 0, 2, 4, 8, 12 and 24 hours post SARS-CoV-2 infection [15]. Using this data, we obtained 543 significantly phosphorylated proteins across all time points profiled (student t-test p-value<0.05 and log2fold change>0.4). Please refer to Supplementary Table S5 for the detailed list of phosphorylated proteins.

RNAi or CRISPR knockdown data

We searched the literature for a list of RNAi or CRISPR screen experiments which test the impact of gene knockdown/knockout on coronavirus load. In particular, we collected RNAi screening data for 5 different coronaviruses: IBV-CoV, MERS-CoV, MHV-CoV, SARS-CoV and SARS-CoV-2 [19, 20, 21, 22, 23, 24, 25, 26, 27, 28]. Additionally, we collected CRISPR screens for SARS-CoV-2 [29, 30]. The combined set of RNAi and CRISPR screen hits is comprised of 482 human proteins in cells infected with a coronavirus, 40 of which were present in our initial SARS-CoV-2 and human network. Table 4 summarizes the hits used in this study while Supplementary Table S6 provides the full list of screen hits used in this study.

General Protein-protein and protein-DNA interactions

Protein-DNA interactions were obtained from our previous work [31], which contains 59578 protein-DNA interactions for 399 Transcription Factors (TFs). Protein-protein interactions (PPIs) were obtained from the HIPPIE database [32], which contains more than 270000 annotated PPIs and for each provides a confidence score which was further used in our network analysis (see below).

Time series transcriptomics

We used two longitudinal transcriptomics datasets to learn the regulatory and signaling networks underlying the SARS-CoV-2 infection. The first was a bulk expression dataset in which iPSC derived lung epithelial cells were infected and profiled before and 1 and 4 hours following infection [33]. The expression of SARS-CoV-2 viral proteins were also quantified for this dataset and we were able to obtain expression levels for 11 viral proteins which were further used in our analysis (see below).

In addition to the bulk data we also used a single-cell RNA-seq data on Calu-3 cell line profiled at 0 (mock), 4, 8, 12 hours post SARS-CoV-2 infection. We filtered all the cells with less than 200 expressed genes or with over 40% mitochondrial genes [34]. We also filtered expressed in less than 3 cells or with a very low dispersion (< 0.15).

Underlying condition lung expression data

We collected and analyzed lung expression data from several reported underlying conditions that impact SARS-CoV-2 infection. Table 1 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 [35], hypertension [36], Diabetes [37], Chronic Obstructive Pulmonary Disease (COPD) [38], and smoking [39]. We also used single cell RNA-seq lung data for gender expression analysis and for inferring aging related genes from [40]. For each dataset, the differentially expressed genes are extracted using the R package 'limma' for microarray data [41] and by using a ranksum test for single cell data.

11

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) [12] method. SDREM integrates time-series bulk 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 [42] 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.

To extend the SDREM to the single-cell level, instead of using DREM as for analyzing bulk transcriptomic data, here we incorporated our previously developed methods SCDIFF and CSHMM [31, 43] to reconstruct the regulatory network underlying the single-cell time-series RNA-seq data. Based on the single-cell regulatory network inference, we generate a list of transcription factors with 3 metrics to evaluate their importance. First, percentage of regulated cells; a TF that regulates a higher percentage of cells will also be assigned with an importance score. Second, p-value for the TF regulation; we calculated a hypergeometric test p-value for each of the predicted TFs based on their target genes. If the target genes of a TF are differentially expressed between points that the TF is regulating, such TF-regulation would be considered more reliable and it should be weighted with a higher score. Namely, TFs will a smaller p-value will be assigned with a higher importance score. Third, we also evaluate the TFs based on their Lasso logistic coefficients from the single-cell regulatory inference (by SCDIFF); a TF with a higher coefficient should be treated with a higher importance. Finally, we calculate an overall importance score IS for each of the TFs predicted from the single-cell regulatory network inference method based on the above 3 scoring metrics.

$$IS(x) = W_{pe} * Score_{pe}(x) + W_{pv} * Score_{pv}(x) + W_{co} * Score_{co}(x) \qquad (1)$$

$$Score_{pe}(x) = (pe(x) - \mu(pe))/\sigma(pe)$$

$$Score_{co}(x) = (co(x) - \mu(co))/\sigma(co)$$

$$Score_{pv}(x) = (pv(x) - \mu(pv))/\sigma(pv)$$

$$pv(x) = -10 * log_2(pvalue(x))$$

12

Where $pe(x), co(x), pvalue(x)$ are the percentage score, coefficient score, and p-value score for TF x respectively. W_{pe}, W_{pv}, W_{co} are weights for each of the scoring metrics. By default, they are all set as 1 for equal importance. However, users are allowed to specify a different set of weights to emphasize specific scoring metrics. If a TF is found to regulate multiple edges of the reconstructed trajectory and thus have multiple overall scores, the maximal one will be used. We ranked all the predicted TFs based on their overall scores. The top ones (under a specific cutoff parameter specified by SDREM method) will be chosen as the final predicted regulators for the following analyses.

The second part of SDREM uses a network orientation method [11], 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.

Using human SARS-CoV-2 transcriptomics and protein phosphorylation data with SDREM

We modified SDREM to improve its performance on the SARS-CoV-2 transcriptomics data. To remove the potential batch effect, we have performed cross-sample normalization between the bulk RNA-seq sample between different time points (control, 1dpi, 4dpi). 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. We integrate the protein phosphorylation data by adjusting the prior for significantly phosphorylated proteins. We first set the prior of all proteins as a default value (e.g, 0.5). Then we scaled the log2fc of the significantly phosphorylated proteins to the range of [0.5, 1] with a min-max normalization method. To mitigate the impact of the outliers, we replaced the min-max values with

5% and 95% percentiles. Proteins (nodes) that are highly phosphorylated will be assigned with a larger prior and they will be favored in the SDREM network analysis.

Identifying key genes in SDREM reconstructed networks

To rank top genes identified by SDREM, we used the strategy described in [11] 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.2).

$$score_w(A) = \frac{\sum\limits_{t \in T} \frac{\sum\limits_{p \in P(t)} w(p)I(A \in N(p)))}{\sum\limits_{p \in P(t)} w(p)}}{|T|}$$

$$w(p) = \prod_{v \in p} Prior_v \prod_{e \in p} weight_e \tag{2}$$

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, $I(*)$ 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, $w(p)$ is the path weight, which is the product of all node priors $Prior_v$ and edge weights $weight_e$ in the path. The node prior is determined using the protein phosphorylation data (if available). The edge weight is based on the strength for the interaction (e.g., PPI)) between the source and target nodes of the edge.

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.3 for details).

$$P_{AB}^{observed} = score_w(A, B) = \frac{\sum\limits_{t \in T} \frac{\sum\limits_{p \in P(t)} w(p)I(A \notin N(P))I(B \notin N(p))}{\sum\limits_{p \in P(t)} w(p)}}{|T|} \tag{3}$$

The above eqn.3 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 versus unweighted versions of the SDREM network. See Supplementary Table S1 (Meta) for details on what was used in the analysis.

Analyzing underlying condition data

We examined several recent studies to determine underlying condition that impact SARS-CoV-2 mortality and infection rates [7, 8]. 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 1. 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' [41]. 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 S7 for details).

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.

15

Potential treatments for top genes

In a similar fashion to Gordon *et al.* [3], we searched public resources (ChEMBL25 [44], IUPHAR/BPS, Pharos [45] and ZINC [46]), as well as literature in order to identify existing drugs and reagents that directly modulate the candidate genes derived from our network reconstruction and condition-specific analyses. Supplementary Table S3 provides a full list of drugs and reagents targeting the identified candidate genes.

Software and data availability

The single-cell extension of the SDREM model (named scSDREM) is implemented in Python. It's publicly available at `https://github.com/phoenixding/sdremsc`. The bulk and single-cell time-series SARS-CoV-2 viral infection transcriptomics data used in this work are available under the GEO accession number GSE153277 and GSE148729, respectively. The regulatory models (by iDREM) for the bulk and single-cell SARS-CoV-2 transcriptomics data is available at : `http://www.cs.cmu.edu/~jund/sars-cov-2`

Acknowledgements

Author contributions

ZB-J, JD, JL-M and YY designed the research; JD, JL-M and YY developed and implemented the methods; JD, JL-M and YY analyzed data; and JD, JL-M, YY, DNK and ZB-J wrote the manuscript.

Conflict of interest

None.

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] A. Stukalov, V. Girault, V. Grass, V. Bergant, *et al.*, "Multi-level proteomics reveals host-perturbation strategies of sars-cov-2 and sars-cov," *bioRxiv*, 2020.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] J. Chen, E. E. Bardes, B. J. Aronow, and A. G. Jegga, "ToppGene suite for gene list enrichment analysis and candidate gene prioritization," *Nucleic Acids Res*, vol. 37, pp. W305–W311, 05 2009.

[14] R. A. Al-Lami, R. J. Urban, E. Volpi, A. M. Algburi, and J. Baillargeon, "Sex hormones and novel corona virus infectious disease (covid-19)," in *Mayo Clin Proc*, Elsevier, 2020.

[15] M. Bouhaddou, D. Memon, B. Meyer, K. M. White, V. V. Rezelj, M. C. Marrero, B. J. Polacco, J. E. Melnyk, S. Ulferts, R. M. Kaake, *et al.*, "The global phosphorylation landscape of sars-cov-2 infection," *Cell*, vol. 182, no. 3, pp. 685–712, 2020.

[16] Y. Jin, S.-J. Lee, R. D. Minshall, and A. M. Choi, "Caveolin-1: a critical regulator of lung injury," *Am J Physiol Lung Cell Mol Physiol*, vol. 300, no. 2, pp. L151–L160, 2011.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] J. Wei, M. M. Alfajaro, R. E. Hanna, P. C. DeWeirdt, *et al.*, "Genome-wide CRISPR screen reveals host genes that regulate SARS-CoV-2 infection," *bioRxiv*, 2020.

[30] B. E. Heaton, J. D. Trimarco, C. E. Hamele, A. T. Harding, *et al.*, "SRSF protein kinases 1 and 2 are essential host factors for human coronaviruses including SARS-CoV-2," *bioRxiv*, 2020.

[31] 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.

[32] 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.

[33] J. Huang, A. J. Hume, K. M. Abo, R. B. Werder, C. Villacorta-Martin, K.-D. Alysandratos, M. L. Beermann, C. Simone-Roach, J. Lindstrom-Vautrin, J. Olejnik, *et al.*, "Sars-cov-2 infection of pluripotent stem cell-derived human lung alveolar type 2 cells elicits a rapid epithelial-intrinsic inflammatory response," *Cell Stem Cell*, 2020.

[34] E. Wyler, K. Mösbauer, V. Franke, A. Diag, L. T. Gottula, R. Arsie, F. Klironomos, D. Koppstein, S. Ayoub, C. Buccitelli, *et al.*, "Bulk and single-cell gene expression profiling of sars-cov-2 infected human cell lines identifies molecular targets for therapeutic intervention," *bioRxiv*, 2020.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] C. Lin and Z. Bar-Joseph, "Continuous-state hmms for modeling time-series single-cell rna-seq data," *Bioinformatics*, vol. 35, no. 22, pp. 4707–4715, 2019.

[44] D. Mendez, A. Gaulton, A. P. Bento, J. Chambers, M. De Veij, E. Félix, M. P. Magariños, 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.

[45] D.-T. Nguyen, S. Mathias, C. Bologa, S. Brunak, *et al.*, "Pharos: Collating protein information to shed light on the druggable genome," *Nucleic Acids Res*, vol. 45, pp. D995–D1002, 11 2016.

[46] T. Sterling and J. J. Irwin, "ZINC 15 - Ligand discovery for everyone," *J Chem Inf Model*, vol. 55, no. 11, pp. 2324–2337, 2015.

Tables

Condition	Accession number	Data type	References
Smoker	GSE10072	Microarray	[39]
Hypertension	GSE24988	Microarray	[36]
Diabetes	GSE15900	Microarray	[37]
COPD	GSE38974	Microarray	[38]
Cancer	GSE2514	Microarray	[35]
Age	GSE136831	scRNA-seq	[40]
Sex	GSE136831	scRNA-seq	[40]

Table 1: **Summary of literature-derived condition lung expression data**. For each condition, we list the accession number, data type and corresponding references.

Gene name	RNA-seq	SARS-CoV-2 interactor?	RNAi supported effect
PISD	Bu	Y	decreased IBV-CoV replication
POU3F2	Bu	N	decreased IBV-CoV replication
RAB7A	Bu	Y	affected MHV-CoV fusion decreased SARS-CoV-2 load
UBE2I	Bu	N	decreased IBV-CoV replication
VPS39	Bu	Y	affected MHV-CoV fusion
ACVR1	Sc	Y	decreased SARS-CoV replication
CAV2	Sc	Y	decreased IBV-CoV replication
CSNK2A1	Sc	N	decreased SARS-CoV replication
ACVR1B	Both	Y	decreased SARS-CoV replication
ATM	Both	Y	decreased SARS-CoV-2 load
CAV1	Both	Y	decreased IBV-CoV replication
DYNC2H1	Both	Y	decreased MHV expression
EPHA2	Both	Y	decreased SARS-CoV replication
G3BP2	Both	Y	decreased SARS-CoV-2 load
MDH1	Both	Y	decreased IBV replication
RBX1	Both	Y	decreased IBV-CoV replication
SMAD3	Both	N	conferred resistance to SARS-CoV-2
SMAD4	Both	N	conferred resistance to SARS-CoV-2
SMARCA4	Both	N	conferred resistance to SARS-CoV-2

Table 2: **Summary of proteins supported by RNAi or CRISPR screen experiments identified by SDREM from time-series bulk and single-cell SARS-CoV-2 expression data sets**. For each protein, we RNA-seq experimental evidence ('Bu' for bulk, 'Sc' for single-cell, and 'Both' for bulk and single-cell), whether it is a known interactor of a SARS-CoV-2 protein and a brief description of the experimental impact on coronavirus load. Table enumerates each protein previously reported to affect coronavirus load in RNAi or CRISPR screen experiments for the set of proteins identified by SDREM from SARS-CoV-2 data (p values: 6.32e-4 for bulk data; 2.50e-3 for single-cell data). Proteins also listed in the top 1000 protein pairs identified by SDREM from either bulk or single-cell data sets are shown in boldface (p values: 7.70e-3 for bulk data; 8.28e-3 for single-cell data).

Gene name	RNA-seq	SARS-CoV-2 interactor?	Approved drugs?
ATR	Bu	Y	N
BRCA1	Bu	N	N
CSNK2B	Bu	Y	N
HSF1	Bu	N	N
NFE2L2	Bu	Y	N
OAT	Bu	Y	N
PLAU	Bu	Y	Y
UBE2I	Bu	N	N
CREB1	Sc	N	N
CREBBP	Sc	N	N
CSNK2A1	Sc	N	N
CSNK2A2	Sc	Y	N
EGFR	Sc	N	Y
ESRRA	Sc	N	Y
HES1	Sc	N	N
HNF4A	Sc	N	N
HSPA8	Sc	N	N
IRAK1	Sc	Y	Y
JAK2	Sc	Y	Y
NEU1	Sc	Y	N
NR1H3	Sc	N	N
NR3C1	Sc	N	Y
RELA	Sc	N	Y
RXRA	Sc	N	Y
ACAT1	Both	Y	N
ATM	Both	Y	N
BRD2	Both	Y	N
EP300	Both	N	N
ERBB2	Both	Y	Y
JUN	Both	N	N
MYC	Both	N	N
STAT6	Both	Y	N

Table 3: **Summary of identified proteins associated with chemical compounds in ChEMBL25, IUPHAR/BPS Guide to Pharmacology, Pharos, or ZINC**. For each human protein, we show RNA-seq experimental evidence ('Bu' for bulk, 'Sc' for single-cell, and 'Both' for bulk and single-cell), whether or not it is a known SARS-CoV-2 and if there are any approved drugs for each target, respectively.

Coronavirus	Screen hits	References
IBV-CoV	94	[21, 24]
MERS-CoV	20	[25, 28]
MHV-CoV	26	[19, 20, 23]
SARS-CoV	135	[22, 26, 27]
SARS-CoV-2	218	[29, 30]

Table 4: **Summary of literature-derived coronaviruses genome-wide RNA or CRISPR screens**. For each screen, we list associated coronavirus, number of significant screen hits and corresponding references. We use 12 different RNAi or CRISPR 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 and SARS-CoV-2) coronaviruses.

Figures
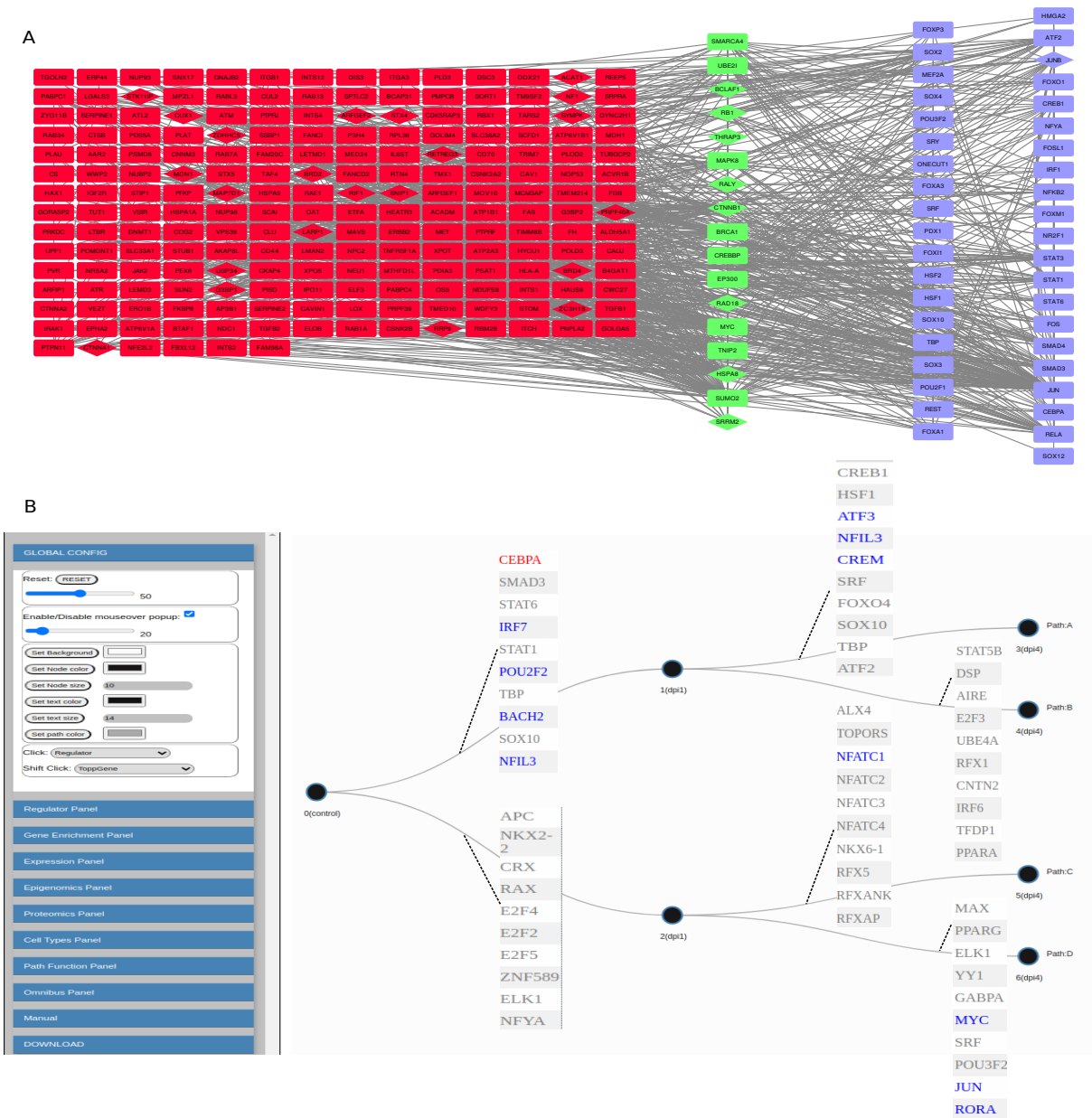
Figures

Figure 1: SDREM predicted signaling network(A) and regulatory network (B) for the time-series RNA-seq dataset (with using protein phosphorylation data). (A) **signaling network reconstructed from the time-series single-cell RNA-seq data**. Red nodes denote sources, green nodes are signaling proteins and blue nodes are TFs associated with regulating the DREM portion of the model. The overlap between the networks is discussed in Results. (B) **The regulatory part of the network** Each path represents a group of genes that share similar expression profiles. The table presented next to each edge indicates the set of transcription factors (TFs) that are predicted to regulate the expression of genes assigned to this path. Red font indicates TF expression is significantly down-regulated, while blue represents up-regulated TFs. TF with stable expression (inferred to be post-transcriptionally or post-translationally regulated) are marked as Gray.

Figure 2: The conserved signaling network between bulk and single-cell SDREM reconstructed networks. Notations in this figure are the same as in Figure 1A.

A — Sum threshold

quantile	0	1	2	3	4	5	6
0.01	0.535492	2.08E-05	0.106374	1	1	1	1
0.05	0.535492	0.000922	0.008137	0.169801	1	1	1
0.1	0.535492	0.00142	0.000842	0.0014	0.007085	0.014895	1
0.15	0.535492	0.016288	4.96E-06	0.000375	0.050248	0.053524	1
0.2	0.535492	0.03398	9.33E-06	9.59E-06	0.145524	0.118562	1
0.25	0.535492	0.037615	2.63E-06	6.50E-06	0.290212	1	1
0.3	0.535492	0.129611	0.000583	7.22E-05	0.174307	1	1

B — Sum threshold

quantile	0	1	2	3	4	5	6
0.01	63	16	1	0	0	0	0
0.05	63	33	6	1	0	0	0
0.1	63	45	13	5	2	1	0
0.15	63	46	23	8	2	1	0
0.2	63	49	27	11	2	1	0
0.25	63	52	33	15	2	0	0
0.3	63	49	30	15	3	0	0

C

gene_symbol	hypertension	COPD	diabetes	smoke	cancer	sex	age	sum
cav1	-1	-1	-1	-1	-1	1	0	-4
jun	-1	1	-1	-1	-1	-1	0	-4
hspa8	-1	-1	0	0	0	-1	0	-3
erbb2	1	-1	0	-1	-1	0	-1	-3
srf	-1	1	-1	-1	-1	0	0	-3
fos	-1	0	1	-1	-1	0	-1	-3
acat1	0	-1	-1	-1	0	-1	1	-3
ctnna1	1	-1	0	-1	-1	-1	0	-3
brd2	-1	1	-1	-1	-1	-1	1	-3
myc	-1	1	0	0	0	-1	-1	-2
akap8l	0	-1	-1	0	0	-1	1	-2
stx4	-1	-1	0	0	0	0	0	-2
creb1	0	-1	1	1	-1	-1	-1	-2
junb	-1	1	0	0	-1	-1	0	-2
crebbp	0	-1	-1	-1	0	0	1	-2
irf1	-1	0	1	-1	-1	1	-1	-2
stat6	0	0	-1	-1	-1	0	1	-2
rela	-1	0	0	0	-1	1	-1	-2
stip1	-1	1	1	1	0	0	0	2
cul2	0	1	0	0	0	0	1	2
atm	1	1	0	1	-1	0	0	2
g3bp1	0	0	-1	1	1	0	1	2
arfgef2	1	0	-1	1	1	-1	1	2
prpf40a	0	0	0	1	1	-1	1	2
bclaf1	0	-1	1	1	0	1	0	2
smad3	0	-1	1	-1	1	1	1	2
larp1	0	0	1	-1	1	0	1	2
sympk	0	0	0	0	1	1	1	3
ep300	0	1	1	-1	0	1	1	3
ptprj	-1	1	1	0	0	1	1	3
csnk2a2	0	1	1	-1	1	0	1	3
zc3h18	0	1	0	0	0	1	1	3
itch	0	1	0	1	0	0	1	3

D

GO biological process complete	raw P value	FDR
positive regulation of gene expression	1.93E-12	3.07E-08
positive regulation of nitrogen compound metabolic process	1.07E-11	5.66E-08
cellular response to chemical stimulus	1.80E-11	5.74E-08
response to stress	1.46E-11	5.81E-08
positive regulation of macromolecule metabolic process	8.77E-12	6.97E-08
cellular response to organic substance	3.54E-11	8.04E-08
positive regulation of cellular metabolic process	3.42E-11	9.05E-08
positive regulation of metabolic process	5.23E-11	9.24E-08
symbiotic process	5.91E-11	9.39E-08
gland development	5.00E-11	9.92E-08
positive regulation of macromolecule biosynthetic process	8.96E-11	1.29E-07
viral process	2.37E-10	2.51E-07
positive regulation of cellular biosynthetic process	2.08E-10	2.54E-07
positive regulation of RNA metabolic process	2.29E-10	2.60E-07
positive regulation of transcription by RNA polymerase II	2.03E-10	2.69E-07
positive regulation of biosynthetic process	2.72E-10	2.70E-07
response to organic substance	3.50E-10	3.27E-07
positive regulation of cellular process	3.75E-10	3.31E-07
positive regulation of biological process	4.41E-10	3.69E-07
positive regulation of transcription, DNA-templated	4.88E-10	3.88E-07
regulation of cellular macromolecule biosynthetic process	7.55E-10	5.72E-07
response to chemical	9.30E-10	6.72E-07

Figure 3: Overlap analysis between condition and SDREM individual gene lists. (A) P values for overlap of the two sets with different quantile and sum score threshold combination for the condition gene set. (B) number of genes in the overlap. (C) 33 selected genes when using a quantile threshold of 0.25 and sum threshold of 2. (D) Significant GO terms for (C).

A

Sum threshold

quantile	0	1	2	3	4	5	6
0.01	0.529557	1.29E-06	0.149318	1	1	1	1
0.05	0.529557	3.74E-05	0.003766	0.234769	1	1	1
0.1	0.529557	0.000337	5.28E-05	0.001174	0.014072	0.021348	1
0.15	0.529557	0.021332	2.95E-06	0.000213	0.09362	0.07605	1
0.2	0.529557	0.012269	6.39E-07	0.000491	0.074106	0.165951	1
0.25	0.529557	0.015355	1.59E-07	1.57E-07	0.45411	1	1
0.3	0.529557	0.094155	3.90E-06	4.05E-05	0.336754	1	1

B

Sum threshold

quantile	0	1	2	3	4	5	6
0.01	91	22	1	0	0	0	0
0.05	91	49	8	1	0	0	0
0.1	91	63	19	6	2	1	0
0.15	91	61	29	10	2	1	0
0.2	91	71	37	12	3	1	0
0.25	91	75	45	21	2	0	0
0.3	91	70	47	19	3	0	0

C

gene_symbol	hypertension	COPD	diabetes	smoke	cancer	sex	age	sum
cav1	-1	-1	-1	-1	-1	1	0	-4
jun	-1	1	-1	-1	-1	0	0	-4
hspa8	-1	-1	0	0	0	-1	0	-3
hax1	-1	-1	0	-1	1	-1	0	-3
erbb2	1	-1	0	-1	-1	0	-1	-3
srf	-1	1	-1	-1	-1	0	0	-3
fos	-1	0	1	-1	-1	0	-1	-3
acat1	0	-1	-1	-1	0	-1	1	-3
ctnna1	1	-1	0	-1	-1	-1	0	-3
brd2	-1	1	-1	-1	-1	-1	1	-3
sympk	0	0	0	0	1	1	1	3
ep300	0	1	1	-1	0	1	1	3
rab1a	0	1	1	-1	1	0	1	3
irak1	0	0	1	-1	1	1	1	3
ptprj	-1	1	1	0	0	1	1	3
neu1	-1	1	1	1	1	-1	1	3
csnk2a2	0	1	1	-1	1	0	1	3
zc3h18	0	1	0	0	0	1	1	3
fgb	-1	1	1	1	1	-1	1	3
plau	0	0	0	1	1	1	0	3
itch	0	1	0	1	0	0	1	3

D

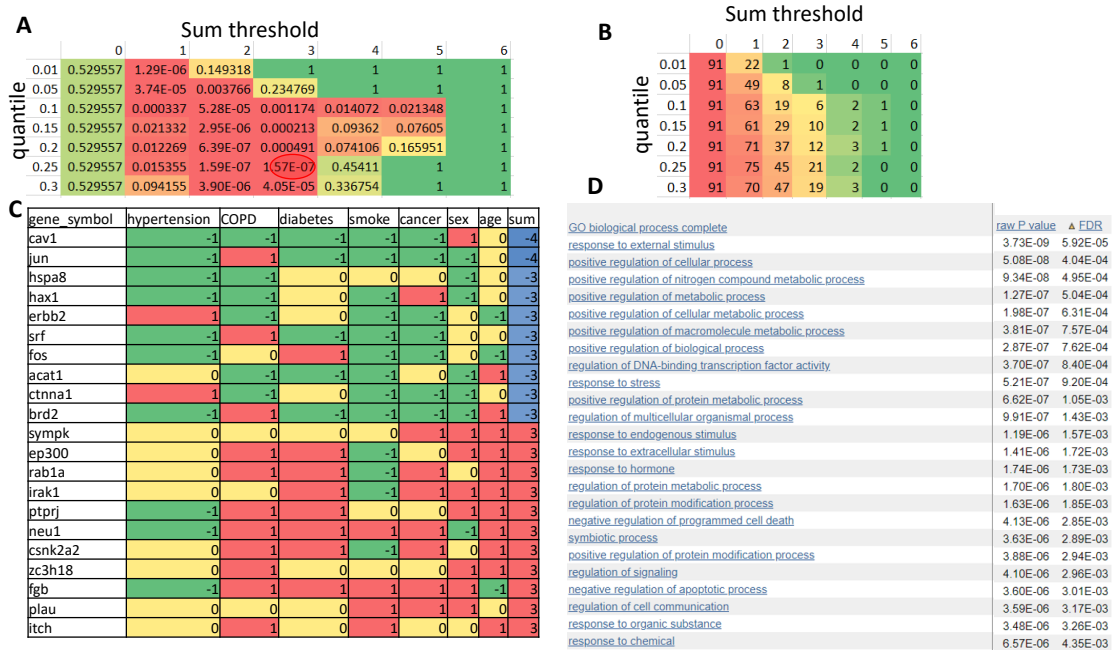| GO biological process complete	raw P value	▲ FDR
response to external stimulus	3.73E-09	5.92E-05
positive regulation of cellular process	5.08E-08	4.04E-04
positive regulation of nitrogen compound metabolic process	9.34E-08	4.95E-04
positive regulation of metabolic process	1.27E-07	5.04E-04
positive regulation of cellular metabolic process	1.98E-07	6.31E-04
positive regulation of macromolecule metabolic process	3.81E-07	7.57E-04
positive regulation of biological process	2.87E-07	7.62E-04
regulation of DNA-binding transcription factor activity	3.70E-07	8.40E-04
response to stress	5.21E-07	9.20E-04
positive regulation of protein metabolic process	6.62E-07	1.05E-03
regulation of multicellular organismal process	9.91E-07	1.43E-03
response to endogenous stimulus	1.19E-06	1.57E-03
response to extracellular stimulus	1.41E-06	1.72E-03
response to hormone	1.74E-06	1.73E-03
regulation of protein metabolic process	1.70E-06	1.80E-03
regulation of protein modification process	1.63E-06	1.85E-03
negative regulation of programmed cell death	4.13E-06	2.85E-03
symbiotic process	3.63E-06	2.89E-03
positive regulation of protein modification process	3.88E-06	2.94E-03
regulation of signaling	4.10E-06	2.96E-03
negative regulation of apoptotic process	3.60E-06	3.01E-03
regulation of cell communication	3.59E-06	3.17E-03
response to organic substance	3.48E-06	3.26E-03
response to chemical	6.57E-06	4.35E-03

Figure 4: Overlap analysis between condition and SDREM gene pair list. (A) P values for overlap genes . (B) number of genes in the overlap. (C) The set of 21 genes when using a quantile threshold of 0.25 and sum threshold of 3. (D) Significant GO terms for (C).