 doi:10.20944/preprints202003.0360.v1

Transcriptional inhibition of host viral entry proteins as a therapeutic strategy for SARS-CoV-2

Xinchen Wang[1*], Ryan S. Dhindsa[1,2], Gundula Povysil[1], Anthony Zoghbi[1,3], Joshua E. Motelow[1,4], Joseph A. Hostyk[1], David B. Goldstein[1,2*]

1. Institute for Genomic Medicine, Columbia University Irving Medical Center, New York, USA
2. Department of Genetics & Development, Columbia University Irving Medical Center, New York, USA
3. Department of Psychiatry, Columbia University Irving Medical Center, New York, USA; New York State Psychiatric Institute, New York, USA
4. Division of Pediatric Critical Care, Department of Pediatrics, New York-Presbyterian Morgan Stanley Children's Hospital, Columbia University Irving Medical Center, New York, USA

*Correspondence: xw2553@cumc.columbia.edu (X.W.), dg2875@cumc.columbia.edu (D.B.G.)

Abstract

There is an urgent need to identify effective therapies for COVID-19 given that a broadly available and effective vaccine is likely at least one year away. Here, we identify compounds that transcriptionally inhibit host proteins required for SARS-CoV-2 entry and should be evaluated for efficacy in SARS-CoV-2 viral infection assays. Recognizing the need for immediately available treatment options, we focused particular attention on FDA-approved drugs that could be immediately repurposed to treat COVID-19 patients. By mining publicly available gene expression data, we identify several compounds that down-regulate *TMPRSS2*, a protein required for SARS-CoV-2 entry that has emerged as a promising therapeutic target. Among these, we find twenty independent studies that implicate estrogen-related and androgen-related compounds as transcriptional modulators of *TMPRSS2* expression, suggesting that these drugs and others acting on the pathway may be promising therapeutic candidates for COVID-19 for further testing. It is also noteworthy that *TMPRSS2* has highly variable and skewed expression in humans, spanning two orders of magnitude with a small minority of individuals having extremely high expression. Combined with literature showing that *TMPRSS2* loss-of-function in mouse is protective against SARS while anti-estrogen treatment predicted to increase *TMPRSS2* expression exacerbates SARS, this observation raises the hypothesis that *TMPRSS2* expression may positively correlate with severity in COVID-19.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 March 2020 doi:10.20944/preprints202003.0360.v1

Introduction

The rapid international spread of the novel pathogenic severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which causes the disease known as COVID-19, poses a global health emergency. As of March 23, 2020, there have been over 332,930 confirmed cases and 14,510 deaths worldwide[1]. The clinical presentation of COVID-19 ranges from mild respiratory symptoms to severe progressive pneumonia, multiorgan failure, and death[2]. Therapeutic interventions beyond supportive care in the literature have included oseltamivir, remdesivir, ganciclovir, α-interferon, hydroxychloroquine and lopinavir[2-7]. Lopinavir, a protease inhibitor, is the only drug with a completed clinical trial but failed to shorten time to improvement or viral shedding. Any effective intervention rapidly mobilized to the frontlines could profoundy impact resource allocation[8]. Effective treatments are therefore vital to handle the surge of COVID-19 infections.

SARS-CoV-2 host factors are attractive targets for therapeutic intervention. The SARS-CoV-2 spike (S) glycoprotein binds the angiotensin-converting enzyme 2 (*ACE2*), allowing the viral particle to enter host cells[9]. Viral entry into host cells also requires cleavage of the viral S protein by host proteases; this cleavage results in irreversible conformational changes to the S protein that allow the virus and host cell membranes to fuse[9]. S protein cleavage, called priming, can use the host serine protease *TMPRSS2* or the cysteine proteases cathepsin B or L (CatB/L)[10-14]. A recent single-cell RNA-sequencing study of human and non-human primate tissues revealed three major cell types that co-express *TMPRSS2* and *ACE2*: type II pneumocytes in the lung, absorptive enterocytes in the terminal ileum, and nasal goblet secretory cells[15]. In addition to the identification of host proteins required for viral entry, a recent preprint mapped other host proteins that interact with 26 of 29 SARS-CoV-2 viral proteins[16].

Computational and *in vitro* screens are useful to identify compounds that either act directly against viral proteins, or that disrupt protein interactions between SARS-CoV-2 and host proteins required for its viral life cycle. Here we propose and develop an approach complementary to approaches by other groups by seeking to identify transcriptional regulators of the host proteins most critical to viral entry and replication within host cells. Given the aggressiveness of this pandemic and the urgency of deploying effective treatments, our first efforts focus on the repurposing of existing drugs as an attractive alternative to novel compound discovery. We note, however, that this screening approach could also be applied to the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 March 2020 doi:10.20944/preprints202003.0360.v1

discovery of new chemical entities with more desirable properties than already available approved medicines.

Here, we use publicly available gene expression profiling data to identify small molecules that down-regulate genes encoding the host proteins required for SARS-CoV-2 viral entry and replication. We primarily focus on identifying candidate expression regulators for all known host proteins required for viral entry, including *ACE2*, *TMPRSS2*, and *CatB/L*. Amongst these candidate targets for therapeutic intervention, *TMPRSS2* appears the most attractive. *TMPRSS2* knockouts in mouse appear normal and do not cause lethality[17], and *TMPRSS2* does not appear to be strongly selected against in the human population (gnomAD probability of loss of function intolerance, pLI = 0, **Supplementary Fig. 1**)[18]. Furthermore, Iwata-Yoshikawa *et al.* demonstrated that after SARS-CoV infection, *Tmprss2* knockout mice have less viral replication in the lungs than wild-type mice, and less severe immunopathology, resulting in milder lung pathology[19].

In contrast, *ACE2* is considered loss-of-function intolerant in humans (pLI = 1, **Supplementary Fig. 1**). In addition, Kuba *et al.* used *Ace2* knockout mice to show that in the process of binding ACE2 for viral entry into host cells, SARS-CoV Spike protein also mediates ACE2 down-regulation[20]. Kuba *et al.* used *Ace2* KO mice to show that this down-regulation contributes to the severity of lung pathologies, thus indicating that down-regulation of *ACE2* may further harm the lungs of infected patients[20]. Finally, *CatB/L* (gene symbols *CTSB*, *CTSL*) are both loss-of-function tolerant in human populations (*CTSB*: pLI = 0; *CTSL*: pLI = 0.01), however *Ctsl* loss in mouse appears to be deleterious. Further, *Ctsb* and *Ctsl* are dispensible for viral entry and spread[14]. While *Ctsb* knockout mice are born normal, without gross abnormalities, and show resistance to induced pancreatitis[21], *Ctsl* knockout mice have hair-loss, skin defects, impaired T cell maturation, dilated cardiomyopathy, and high postnatal mortality[22]. Furthermore, *Ctsl* knockout mice have higher mortality after influenza A infection compared to wildtype which might be due to defective immune responses caused by the *Ctsl* knockout[23]. For these reasons, we focus on transcriptional modulation of *TMPRSS2* as the highest priority, but also include transcriptional modulation results for all four host factor proteins. In addition, we also identify small molecules that down-regulate broad sets of human proteins that form protein-protein interactions with SARS-CoV-2 viral proteins. The initial identification of these compounds highlights potential therapeutic targets and pathways that could pursued by drug repurposing for the amelioratation COVID-19 symptoms in humans.

 doi:10.20944/preprints202003.0360.v1

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 March 2020 doi:10.20944/preprints202003.0360.v1

Results

Literature-wide screen for transcriptional inhibitors of SARS-CoV-2 host factors reveals drug repurposing candidates

To identify compounds that transcriptionally inhibit host factors required for SARS-CoV-2 viral entry, we performed a literature-wide screen of RNA-seq datasets in the NCBI Sequence Read Archive (SRA) that incorporated keywords relating to drug treatments. Of 252,877 human RNA-seq datasets in the SRA that were uniformly mapped by the Skymap project[24], we identified 29,550 samples in 1,222 studies that involved a drug treatment (Methods). Within each study, we manually assigned samples as case or control for each comparison group based on sample descriptions and literature reviews, yielding 3,089 distinct case-control comparisons which we used to create a database of differentially expressed genes under various drug treatments.

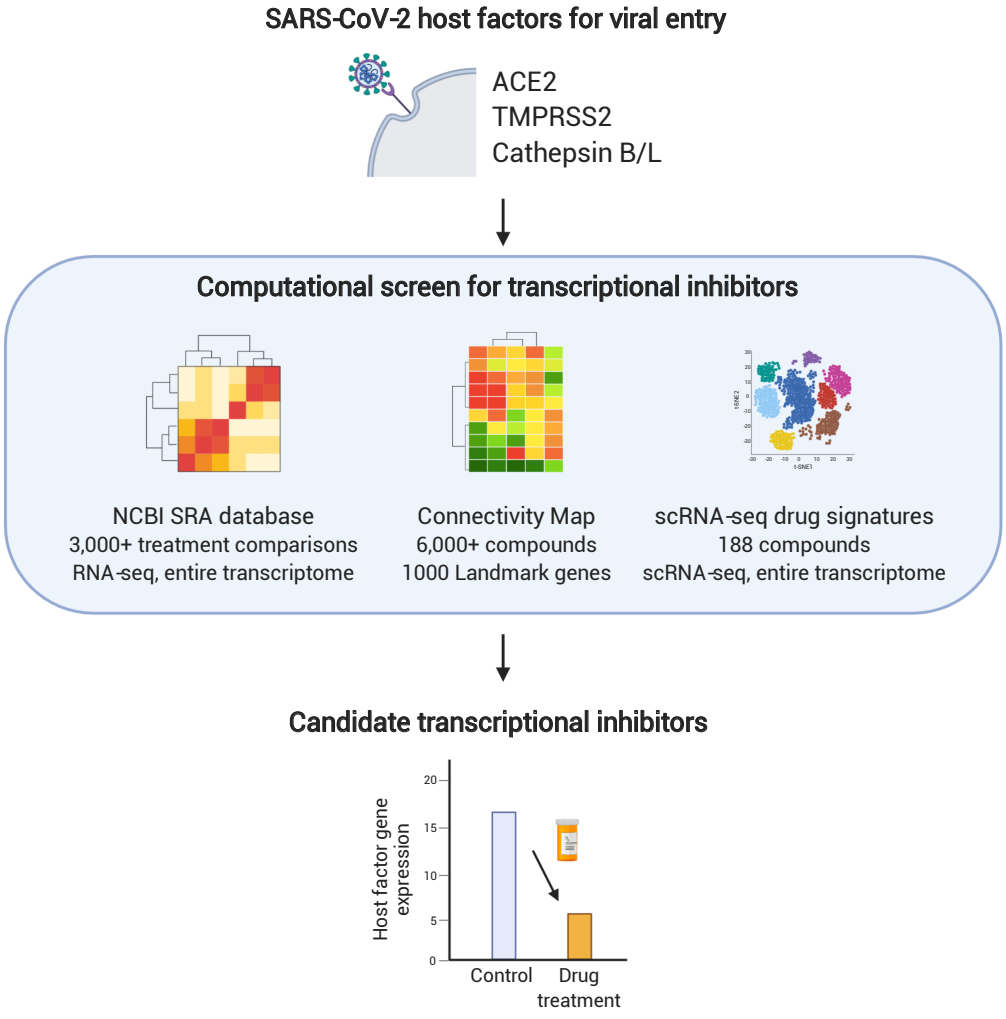


Fig. 1: Overview of repurposing approach to identify transcriptional inhibitors of SARS-CoV-2 host factors

 doi:10.20944/preprints202003.0360.v1

Identification of TMPRSS2 transcriptional inhibitors

Next, we queried this database to identify drug treatments that led SARS-CoV-2 host factor genes to be significantly differentially expressed. We first focused on *TMPRSS2*, given its promise as a candidate for therapeutic intervention. At a Bonferroni-corrected p-value of 0.05 (raw $p < 2.06 \times 10^{-5}$), we identified 32 treatment conditions that led to significant down-regulation of *TMPRSS2* in human cell lines, and 76 conditions that led to up-regulation. While the drugs that down-regulate *TMRPSS2* may be useful as viral targets, the drugs that up-regulate are also important to recognize because they may exacerbate viral infection. Notably, 12 of 32 drug treatments that significantly down-regulated TMPRSS2 (seven independent studies) and 24 of 76 treatments that led to TMPRSS2 up-regulation (15 independent studies) involved estrogens, androgens, or agonists or antagonists or their receptors. These results are consistent with studies showing that TMPRSS2 is regulated by androgens[25]. Specifically, treatment with estradiol (longer than 3 hours, see **Supplementary Fig. 2** for time-course data), genistein (a phytoestrogen that modulates ERα and ERβ), and MDV3100/enzalutamide (an androgen receptor antagonist) led to statistically significant down-regulation of *TMPRSS2* between 1.6-fold and 14-fold, depending on experimental conditions, cell lines and choice of controls (example in **Fig. 2**, data in **Supplementary Table 1**). As *TMPRSS2* is commonly fused with the ETS transcription factor *ERG* in prostate cancer, we also replicated the enzalutamide signal in LAPC4 cells, which are known not to harbor *TMPRSS2-ERG* rearrangements (**Fig. 2**)[26]. Conversely, *TMPRSS2* expression increased between 1.4-fold and 20-fold following treatment with androgens (e.g. testosterone, any duration of treatment, see **Supplementary Fig. 2**), synthetic androgens (R1881/Metribolone) or short-doses of estradiol (under 3 hours, see **Supplementary Fig. 2**).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 March 2020 doi:10.20944/preprints202003.0360.v1

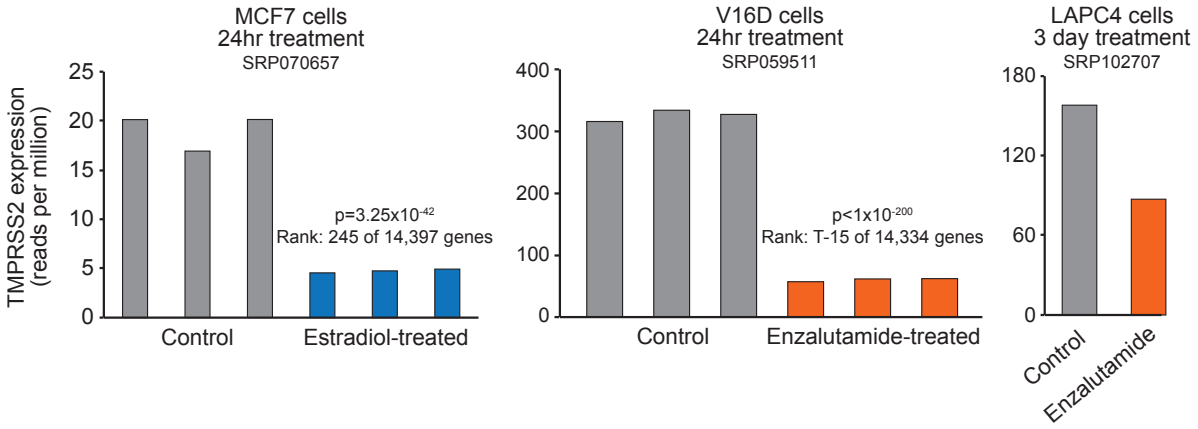


Fig. 2. **Example effects of transcriptional inhibitory effects of estradiol and enzalutamide treatment on TMPRSS2 expression.** Data from RNA-seq studies on MCF7 cells (breast cancer line), V16D cells (prostate cancer line) and LAPC4 cells (prostate cancer line, no known TMPRSS2 fusion). Bars correspond to distinct biological replicates. Difference in control expression of TMPRSS2 corresponds to differences in baseline TMPRSS2 expression in prostate & breast tissue (see Supplementary Fig. 3). p-values calculated using DESeq2, comparing entire transcriptome between treatment and control. Rank represents rank of TMPRSS2 differential expression p-value compared to all other genes in transcriptome. SRP value corresponds to accession number of corresponding study in the NCBI SRA.

In addition to estrogen and androgen-related compounds, other treatments that decrease *TMPRSS2* expression include: dual TGF-beta and EGF treatment (347-fold decrease in expression in HeLa cells, single study only), and chaetocin (non-specific histone lysine methyltransferase inhibitor, ~4-fold decrease).

To identify more compounds that modulate *TMPRSS2* expression, we considered data from the Connectivity Map[27], which includes an unbiased screen of compounds in multiple cell types followed by expression profiling of 978 landmark genes (L1000 platform) and statistical imputation for the rest of the transcriptome. However, *TMPRSS2* was not well-imputed by the Connectivity Map (self-correlation = 0.56 between RNA-seq expression and imputed expression). Consistent with a recent preprint on the lack of reproducibility for L1000-based gene imputation values[28], we did not observe a significant difference in *TMPRSS2* expression following estradiol treatment in MCF7 breast cancer cells or PC3 prostate cancer cells (breast cancer and prostate cancer cells were used in RNA-seq studies from above).

We also considered single-cell RNA-sequencing (scRNA-seq) data from a recent study that measured transcriptomic signatures following treatment with 188 compounds in three cancer cell lines (MCF7 breast cancer, K562 leukemia, A549 lung adenocarcinoma) [29]. *TMPRSS2* expression was only present in MCF7 cells, and in MCF7 we identified two

compounds that led to statistically significant increase in *TMPRSS2* expression: JQ-1, a BET bromodomain inhibitor (q-value = $1.58x10^{-24}$, normalized effect size = 0.19), and fulvestrant, an estrogen receptor antagonist (q-value = $3.96x10^{-11}$, normalized effect size = 0.14). Together, these results identify existing drug compounds that can potentially be repurposed to transcriptionally inhibit *TMPRSS2* expression, and suggest that the activation of estrogen pathways or inhibition of androgen pathways can be a promising modality for clinical intervention in SARS-CoV-2 infection.

Identification of ACE2 transcriptional inhibitors

We next searched for compounds that transcriptionally modulate *ACE2* expression, however we note that while *ACE2* is used as a host factor for viral entry, virus-induced loss of *ACE2* expression is believed to exacerbate SARS-CoV symptoms. *ACE2* shows a very tissue-restricted expression pattern, reducing the number of experiments in which comparisons could be performed (**Supplementary Fig. 3**). Despite this limitation, we identified 9 comparison conditions that led to ACE2 down-regulation, and 24 that lead to *ACE2* up-regulation (**Supplementary Table 2**). Within these nine comparisons, we noticed that two BET bromodomain inhibitors, JQ-1 and CPI-203, led to reduction of *ACE2* expression. Notably, the CPI-203 treatment comparison was performed in bronchial epithelial cells, and treatment led to ~6-fold *ACE2* down-regulation in cells from both healthy patients and cystic fibrosis patients (38th strongest change out of 15,701 genes tested upon CPI-203 treatment, within-study FDR = $2.75x10^{-20}$, **Supplementary Fig. 4**)[30]. A search in our database for other treatments involving BET inhibitors comparisons yielded two additional comparisons where *ACE2* differential expression was tested (in A375 and SET2 cells), however both comparisons were not statistically significant, likely due to low baseline expression of *ACE2* in those cell lines. Notably, the synthetic androgen R1881 was among compounds that up-regulate *ACE2*, in addition to *EGFR* inhibitors (gefitinib, erlotinib, WZ4002). *ACE2* expression was poorly imputed by the Connectivity Map (self-correlation = 0.38), and is poorly expressed in the three cell lines profiled by the drug transcriptome RNA-seq study.

Identification of CTSB and CTSL transcriptional inhibitors

Finally, we considered targets that could lead to transcriptional inhibition of cathepsin B and cathepsin L. 44 treatment conditions led to decreases in *CTSB* expression, and 74 led to creases in *CTSL* expression. Notably, cardiac glycosides used for heart failure such as proscillaridin and digoxin led to decreases in both *CTSB* and *CTSL* expression (~4-8-fold

 doi:10.20944/preprints202003.0360.v1

decreases in expression, **Supplementary Tables 3** and **4**). In the Connectivity Map, expression of *CTSL* was directly assayed by the L1000 array, and expression of *CTSB* was imputed to a higher accuracy than both *TMPRSS2* and *ACE2* (self-correlation = 0.82), suggesting Connectivity Map data could be used for compound identification for both CTSB and CTSL transcriptional inhibition (outputs available in **Supplementary Tables 5** and **6**). For both genes, the Connectivity Map analysis identified over 100 compounds that led to either significant up-regulation or down-regulation of *CTSB* or *CTSL*, and of note, treatment with the cardiac glycoside digoxin led to down-regulation of both cathepsin genes across a range of dosages.

Broad transcriptional inhibitors of host proteins that interact with SARS-CoV-2 viral proteins

We next considered a larger set of 332 host proteins that may be required for viral infection based on their protein-protein interactions with SARS-CoV-2 viral proteins[16]. These proteins were recently identified by affinity-purification mass spectrometry of SARS-CoV-2 proteins expressed in human cells. Using the Connectivity Map, we sought to identify compounds that could down-regulate these host proteins. The Connectivity Map uses a “connectivity score” to assess each tested compound’s ability to reverse a query signature. This score ranges from -100 to +100, with a score of -100 indicating complete reversal. We used three different subsets of the 332 host genes. First, we only considered the 33 genes included in the 978 landmark genes directly profiled in the L1000 assay (**Supplementary Fig. 5A**). In the second query, we added in an additional 33 genes that are well-imputed (**Supplementary Fig. 5B**). In the third query, we included the top 150 genes by fold-change in affinity-purification from the protein-protein-interaction map, regardless of imputation quality (**Supplementary Fig. 5C**). In total, we identified 12 compounds that achieved a Connectivity Score stronger than the recommended cutoff of -90 across all three queries (**Supplementary Table 7**). Because these compounds target more proteins than only those required for viral entry, these are candidates to be efficacious in more broadly limiting viral entry and replication.

Expression patterns of TMPRSS2 in human populations

Recent epidemiological data indicate that COVID-19 may cause severe illness in up to 30% of infected individuals with a case fatality-rate exceeding 20% in high risk populations[2,31]. In children, the data are more limited, but the clinical presentation is milder including a substantial asymptomatic carrier rate and a case fatality-rate < 1%[32,33]. Sex-specific infection and severity rates vary, with some data pointing to equal infection rates between men and women, and other

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 March 2020 doi:10.20944/preprints202003.0360.v1

data showing higher rates of infection and critical illness in men[2,34-36]. Given that *TMPRSS2* expression is modulated by the sex hormones estradiol and androgens, and given evidence that *TMPRSS2* loss of function in mouse could be protective against severity of SARS-CoV and other viruses in mouse and human[19], we next asked whether demographic differences in expression of *TMPRSS2*, as well as *ACE2*, could explain trends observed in COVID-19 epidemiological data.

The most striking feature of *TMPRSS2* and *ACE2* expression data is how variable its gene expression is amongst individuals. Using post-mortem gene expression data from human lungs (n=427, age range 20-80, GTEx Consortium v7)[37], we noticed that both *TMPRSS2* and *ACE2* have highly variable expression amongst individuals (**Fig. 3**). This variability is not due to differences between age groups or sex, but is present within each demographic group. As an example, we considered a set of 156 men aged 40-59. In this group, *TMPRSS2* expression in the lung varied by two orders of magnitude from 2.3 TPM (transcripts per million) to 249.5 TPM (median 44.3 +/- 32.0 TPM). Further, while median *ACE2* expression is much lower than *TMPRSS2* (median TPM 1.1 in same individuals), we observed a similarly broad spread in expression spanning two orders of magnitude, with *ACE2* expression ranging from 0.17 TPM to 14.5 TPM in the same demographic group. Given the high spread in the expression of *TMPRSS2* and *ACE2*, as well as literature evidence that loss of *TMPRSS2* is associated with severity of infection[19], we hypothesized that the high variance in expression of *TMPRSS2* and *ACE2* could be associated with the wide range in COVID-19 severity observed in the human population. To test this, we quantified the skewness in expression of the four SARS-CoV-2 host factors, and noticed that all four genes have highly skewed expression patterns (skewness between 1.6 and 5.6), where the majority of individuals tested have low expression of each gene, but a small subset of individuals have very high expression. To quantify whether this is unusual, we compared the expression skewness of these four genes against every other gene expressed in human lung samples (minimum expression > 1 TPM). Notably, the expression of all four genes is more skewed than the median gene expressed in lung (69th percentile or above for all genes, **Fig. 4A**). *ACE2* in particular has one of the most skewed expression patterns in the entire transcriptome (97th percentile), and even among genes with similar expression levels (94th percentile among genes with median TPM between 0.5 and 3, **Fig. 4B**). These results are consistent with a model where a small subset of individuals have high expression of SARS-CoV-2 host factors and are at particularly high risk for being infected.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 March 2020 doi:10.20944/preprints202003.0360.v1

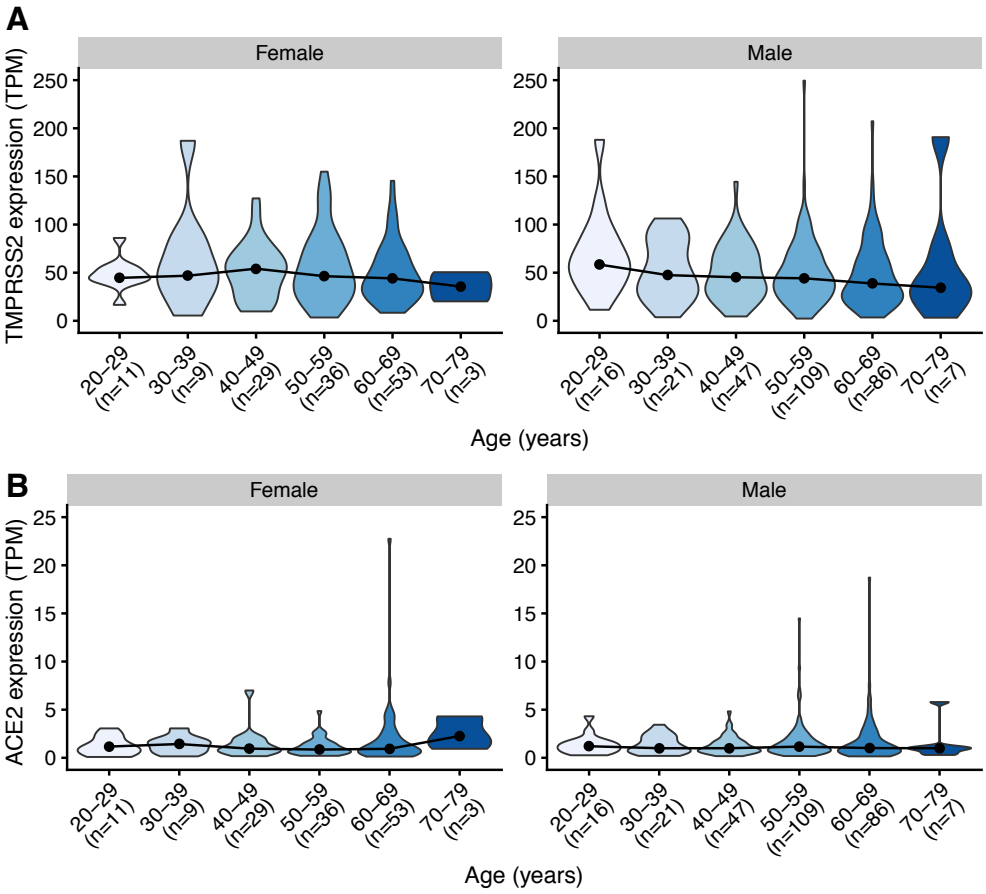


Fig. 3. **Distribution of gene expression for TMPRSS2 and ACE2 in human lungs.** Data from GTEx consortium (v7), samples split by sex and age group. Expression values represented in transcripts per million (TPM)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 March 2020 doi:10.20944/preprints202003.0360.v1

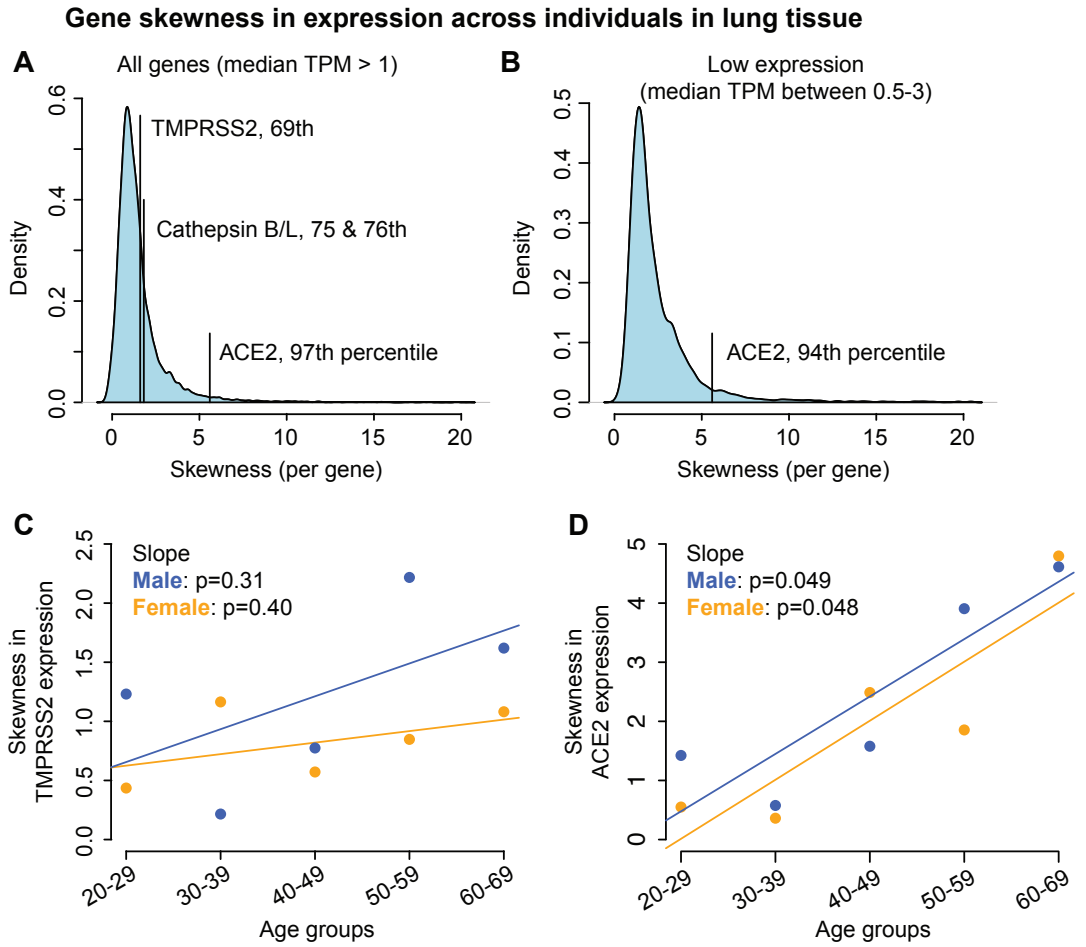


Fig. 4. Skewness in gene expression of SARS-CoV-2 host co-factors. (A) Skewness in gene expression across 427 lung samples for TMPRSS2, Cathepsin B/L and ACE2. Numbers after gene names correspond to percentiles of skewness compared to entire transcriptome. (B) ACE2 has high skewness even when compared to genes matched for low expression values. (C,D) Skewness in expression of TMPRSS2 (panel C) and ACE2 as a function of age (panel D). p-values calculated from linear regression of skewness against age.

In addition, we considered sex and age-specific gene expression patterns for the four host factor genes. Surprisingly, we did not observe strong sex-specific differences in the expression of any of the four genes, and the expression of all four genes is relativity consistent across the GTEx samples aged 20-80 (**Fig. 3, Supplementary Fig. 6**). However, we did notice that the skewness of expression for *ACE2* increased significantly with age (p=0.048 in females, p=0.049 in males, **Fig. 4D**). A more modest but not statistically significant increase in skewness was observed for *TMPRSS2* expression across age (**Fig. 4C**). As epidemiological data suggests that children are less severely affected by the SARS-CoV2 infection[32], we also investigated the expression patterns of *TMPRSS2* and *ACE2* in infants and children. We used RNA-seq data

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 March 2020 doi:10.20944/preprints202003.0360.v1

from sorted lung cells compiled by the LungMap consortium[38], which includes lung RNA-seq data from infants, children and adults. Both *TMPRSS2* and *ACE2* have the highest expression in infants, but while one of two RNA-seq datasets shows a slight increase in *TMPRSS2* expression from childhood to adult, there is no clear trend indicating that adults have much higher expression of these two genes (**Supplementary Fig. 7**).

Relevance in lung tissue of estradiol-based transcriptional inhibition of TMPRSS2

One key limitation of the drug repurposing analysis presented above is that all experiments were performed in breast cancer or prostate cancer cell lines, leading to concerns about applicability and relevance to lung tissue. We next asked whether treatment with estradiol or androgen receptor antagonists could potentially transcriptionally inhibit *TMPRSS2* expression in human lung tissue. First, we note that in mouse, androgen treatment in castrated male mice was shown to significantly increase *Tmprss2* expression in the lung[25]. To test whether this would extend to human lungs, we leveraged *TMPRSS2*'s extremely high variability in gene expression across individuals. Specifically, we hypothesized that if *TMPRSS2* expression in human lung tissue were regulated by estrogen and androgens, then its expression in lung would vary in tandem with known estrogen and androgen response genes (**Fig. 5A**). To test this hypothesis, we quantified the correlation in gene expression between *TMPRSS2* and 12,276 other genes expressed in lung across all 427 individuals with RNA-seq data (TPM > 5, Spearman correlation, GTEx v7 data). Notably, using Gene Set Enrichment Analysis (GSEA) on the gene list ranked by correlation with *TMPRSS2* expression[39], we observed that the hallmark early and late estrogen response gene sets, as well as the hallmark androgen response gene set are three of the top four sets most enriched in genes with high positive correlation *TMPRSS2* expression in lung (**Fig. 5B,C**).

Finally, we tested whether the transcriptionally repressive direction of effect for TMPRSS2 modulation by estradiol and androgen receptor antagonists might be the same in the human lung as it is in cell lines from **Fig. 2**. To do so, we considered temporal gene expression data from MCF7 breast cancer cells and LNCaP prostate cancer cells in response to estradiol and dihydrotestosterone (DHT, androgen) treatment (SRA accession #: SRP070657 and SRP059762) [40]. Notably, we observed that in human post-mortem lungs, genes with expression patterns that most strongly correlate with *TMPRSS2* also behave similarly to *TMPRSS2* in response to estradiol and DHT treatment in MCF7 and LNCaP cells, respectively (**Fig. 5D,** middle panels). Conversely, genes with expression patterns most inversely correlated with

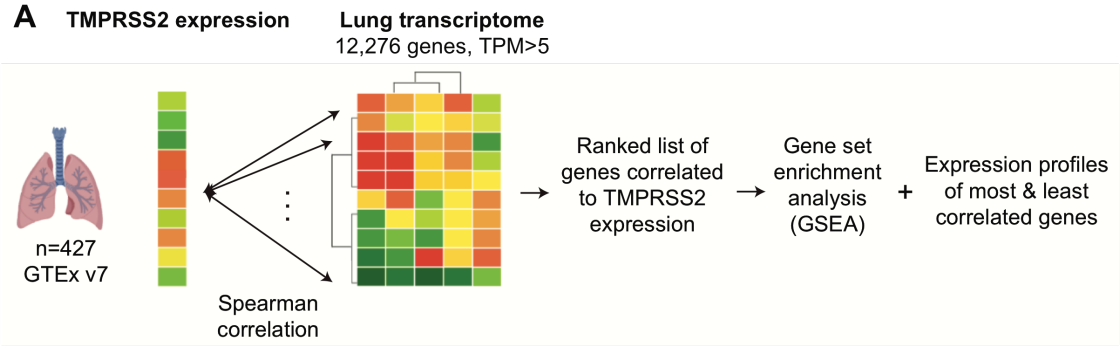
 doi:10.20944/preprints202003.0360.v1

TMPRSS2 in human lungs behave inversely in response to estradiol and DHT treatment (**Fig. 5D,** right panels). Collectively, these results indicate that expression of *TMPRSS2* in the human lung changes alongside known estrogen and androgen response genes, and with the same direction of effect. This suggests that the expression of *TMPRSS2* in the lung can be repressed by treatment with estrogens or androgen receptor antagonists, and the expression can be activated by androgens. Furthermore, these data suggest that within human populations, TMPRSS2 expression in the lung is modulated by estrogens and androgens.

(next page) Fig. 5. Relevance of estrogen-based TMPRSS transcriptional inhibition in lung. (A) Overview of approach to test whether estrogen and androgen-response genes correlate with TMPRSS2 expression in lung (n=427). (B) Top five enriched gene sets reveals expression of estrogen and androgen response genes in lung are highly correlated with expression of TMPRSS2. (C) Gene set enrichment plots for estrogen and androgen response sets. (D) Genes with most correlated expression to TMPRSS2 in lung also have the same estradiol and DHT-dependent gene expression patterns in MCF7 and LNCaP cells, respectively. Genes with least correlated expression to TMPRSS2 in lung have estradiol and DHT-dependent expression patterns different from TMPRSS2. Shared areas correspond to 95% CI.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 March 2020 doi:10.20944/preprints202003.0360.v1

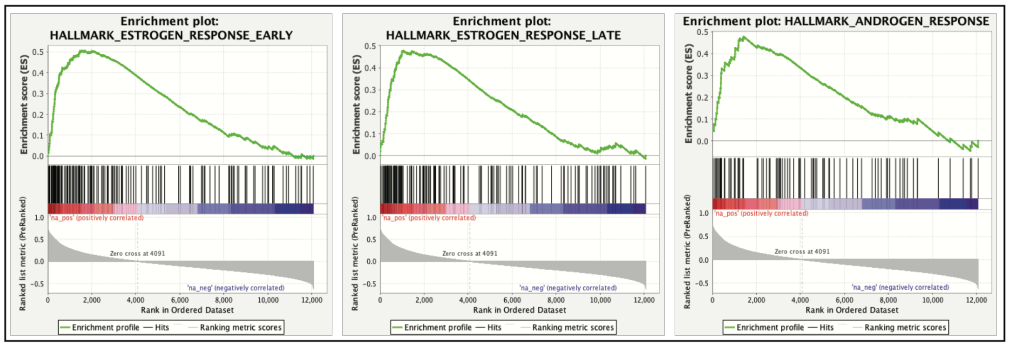
A



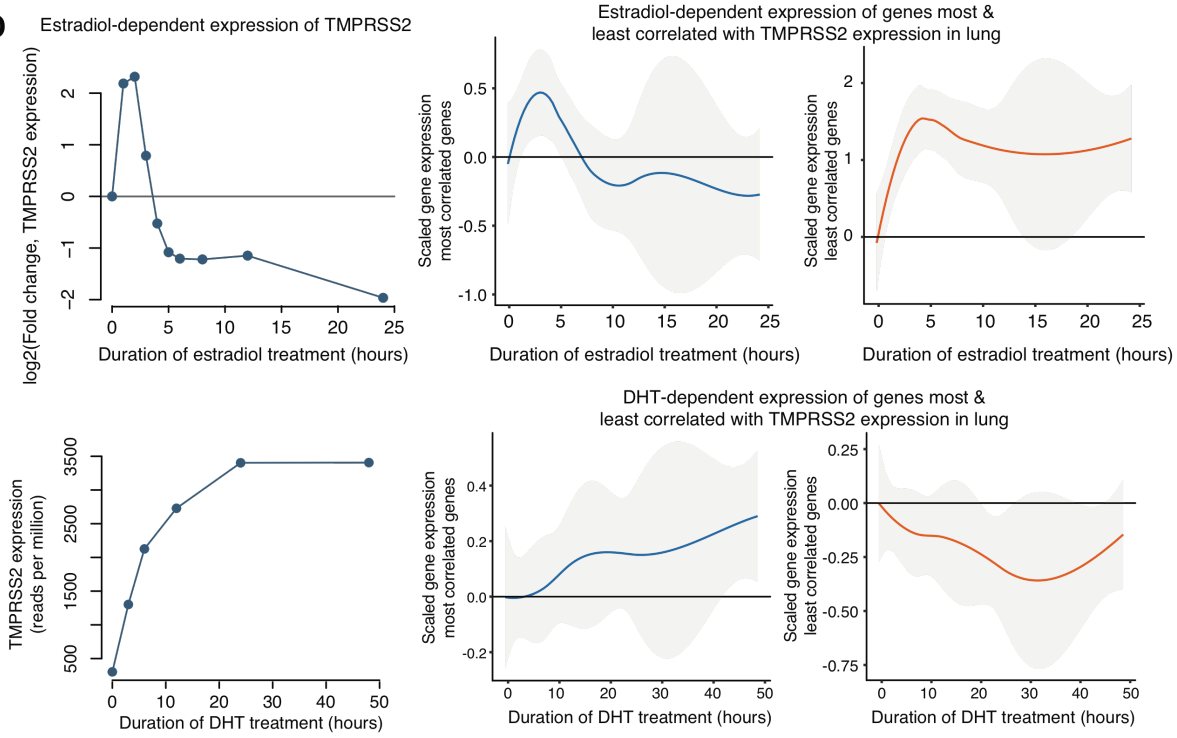
B

Gene Set	Size	Normalized enrichment score	FDR
Hallmark: Cholesterol homeostasis	70	2.81	0.000
Hallmark: Estrogen response early	156	2.70	0.000
Hallmark: Estrogen response late	147	2.57	0.000
Hallmark: Androgen response	88	2.34	0.000
Hallmark: mTORC1 signaling	179	2.06	0.001

C



D



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 March 2020 doi:10.20944/preprints202003.0360.v1

Discussion

Here, we have used a computational approach that leverages publicly available transcriptomic data investigating how a large number of compounds, including hundreds of FDA-approved drugs, perturb gene expression in multiple cell types. These analyses have identified a number of FDA approved drugs that have been shown to reliably down-regulate the host proteins that are critical to the entry of SARS-CoV-2 into host cells. Amongst these, the estrogen-related compounds estradiol, genistein, and the androgen receptor antagonist enzalutamide are shown to downregulate *TMPRSS2*, which is required for SARS-CoV-2 spike protein priming, and appear to be the most promising repurposing candidates for symptom amelioration in COVID-19 patients. Convergence of multiple compounds on the same biological pathway suggests that other commonly prescribed drugs acting on the same pathway but not within our database may also be promising repurposing candidates to transcriptionally inhibit *TMPRSS2*, including dutasteride and finasteride, which inhibit DHT production, and oral contraceptive pills that contain estrogen. Additionally, a common intervention such as the use of spironolactone to ameliorate potassium loss due to loop diuretics in the treatment of acute respiratory distress syndrome may provide unforeseen benefits given its antiandrogenic effect[41,42]. Consistent with these data, a mouse study of SARS-CoV infection showed that in female mice, ovariectomy or treatment with estrogen receptor antagonists (both expected to increase lung *Tmprss2* expression) resulted in increased mortality to SARS-CoV infection[43]. Of the host factors involved in SARS-CoV-2 entry, *TMPRSS2* appears the most promising candidate for transcriptional inhibition because: (i) *TMPRSS2* expression levels have been shown to be associated with severity in mouse models, (ii) population genetic considerations suggest that transcriptional inhibition of *TMPRSS2* may not be harmful, and (iii) *TMPRSS2* is essential for SARS-CoV-2 entry to host cells, unlike cathepsins B and L. Finally, we emphasize that the core hypotheses underlying the transcriptional inhibition approach are testable with observational studies to assess whether widely-used drugs affecting estrogen and androgen signaling correlate with COVID-19 severity.

On the other hand, there are some indications that although down-regulation of ACE2 may reduce coronavirus transmissibility, it may also contribute to virulence[44]. Moreover, population genetics data suggest that downregulation of ACE2 is likely to be damaging, at least when genetically mediated[18]. Meanwhile, *CatB/L* does not appear to be essential for viral entry[11]. While it is already known that sex hormones can regulate *TMRPSS2* expression, this work shows estrogen-related compounds and androgen receptor antagonists appear to be the most

 doi:10.20944/preprints202003.0360.v1

securely identified down-regulators of TMPRSS2 expression amongst FDA approved drugs and other widely tested compounds. It is therefore a high priority to evaluate how estrogen-related compounds perform both in *in vitro* viral entry assays and in symptom amelioration in patients. While lower priority, the regulators of *ACE2* and *CatB/L* should also be evaluated in *in vitro* viral entry assays, and it seems possible that *ACE2* down-regulation could theoretically play a role in prophylactic prevention of infection, but not reduction in symptom severity, if such transcriptional inhibition were shown to be safe.

Of particular interest is the wide variability of *TMPRSS2* expression levels in the human population, suggesting a possible explanation for much of the variability amongst people in the severity of disease. While we do not observe a strong correlation between increased *TMPRSS2* or *ACE2* expression and age in adults, we do observe a significant increase in skewness of *ACE2* expression with age and non-significant increase in skewness of *TMPRSS2*, which raises a possible connection between higher expression of these host factors and the marked increase in mortality rate in the older population. Further, the dramatic variation in expression of both *TMPRSS2* and *ACE2* also suggests a credible hypothesis for variation of vulnerability within age groups. We note that this hypothesis can be readily tested by evaluating *TMPRSS2* expression levels in patients with a range of severities, including comparisons of all symptomatic patients with the general populations. We consider these assessments a high priority, and suggest consideration of whether these evaluations can be performed in the same RNA samples used for SARS-CoV-2 diagnostics, given that scRNA-seq studies have shown *TMPRSS2* and *ACE2* expression in nasal goblet cells.

In summary, we have developed a therapeutic strategy designed to complement existing antiviral strategies focused on inhibition of viral proteins and strategies focused on small molecule disruption of host protein interactions with the virus. Our strategy seeks to transcriptionally inhibit the key proteins that SARS-CoV-2 relies upon, and has identified immediate opportunities for therapeutic intervention using approved estrogen-related compounds or anti-androgens that modulate expression of *TMPRSS2*, critical to viral entry. Depending on the degree of *TMPRSS2* transcriptional inhibition needed for a protective effect, we hypothesize these transcriptional inhibitors can be used either alone or in combination with other direct inhibitors. Further, this framework can be expanded to identify the most effective down regulators of both viral entry proteins and proteins critical to the life cycle of the virus within cells.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 March 2020 doi:10.20944/preprints202003.0360.v1

Supplementary Tables

Supplementary Table 1: Transcriptional modulators of TMPRSS2 by RNA-seq using NCBI SRA

Supplementary Table 2: Transcriptional modulators of ACE2 by RNA-seq using NCBI SRA

Supplementary Table 3: Transcriptional modulators of CTSB by RNA-seq using NCBI SRA

Supplementary Table 4: Transcriptional modulators of CTSL by RNA-seq using NCBI SRA

Supplementary Table 5: Transcriptional modulators of CTSB by Connectivity Map

Supplementary Table 6: Transcriptional modulators of CTSL by Connectivity Map

Supplementary Table 7: Transcriptional inhibitors of host proteins with protein-protein interactions to SARS-CoV-2 viral proteins

Acknowledgements

We wish to thank David D. Pollock, Nasa Sinnott-Armstrong and Sahin Naqvi for very helpful comments on the manuscript, and Kunal Bhutani for helpful discussions on parsing the Sequence Read Archive.

Declaration of Interests

D.B.G. is a founder of and holds equity in Pairnomix and Praxis, serves as a consultant to AstraZeneca and has received research support from Janssen, Gilead, Biogen, AstraZeneca and UCB.

Methods

Computational screening for transcriptional inhibitors

We downloaded uniformly processed RNA-seq datasets from 3,764,506 high-throughput sequencing samples from SkyMap in raw count format[24]. We subsetted these for RNA-seq datasets performed on human samples, yielding 252,877 samples. Potential studies incorporating drug treatments were identified by searching for studies incorporating the term “treatment” or “nM”, yielding 29,550 unique samples (designated by SRS ID) in 1,222 unique studies (designated by SRP ID). Samples within each study were manually assigned to case or control status, matching for other conditions (e.g. cell line used). To ensure consistency, all manual assignments were performed by a single Ph.D-level researcher with expertise in wet-lab experimental design using sample annotations present in the SRA, as well as reading relevant papers when case-control assignments are ambiguous or insufficient annotations were provided. Samples were allowed to be assigned to multiple comparisons (e.g. same samples used as controls for different drug perturbations), and samples with similar conditions (e.g. similar dosages or treatment times) were grouped to increase statistical power in some cases. In these cases, non-grouped sample comparisons were also kept as a separate comparison entry. In total, the 1,222 studies yielded 3,089 case control comparisons. Within each comparison, differentially expressed genes were identified using the DESeq2 R package on raw

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 March 2020 doi:10.20944/preprints202003.0360.v1

count data, and for the current analysis, we focused specifically on a subset of comparisons with at least one biological replicate. scRNA-seq drug treatment data was downloaded from Srivatsan *et al.*[29].

TMPRSS2 gene expression analyses

TMPRSS2 gene expression across human individuals were downloaded from the GTEx project (version 7). Correlation between TMPRSS2 and other genes expressed in lung (n=427) was performed in R by calculating the Spearman correlation between TMPRSS2 expression and other genes, restricting to only 12,276 genes with median expression greater than 5 transcripts per million. Skewness of all expression patterns was calculated using the “skewness” function in the e1071 package in R.

Gene Set Enrichment Analysis was run using GSEA (v4.0.3) using the pre-ranked function, with values for each gene from -1 to 1 taken from their Spearman correlation with TMPRSS2 expression in human lung samples.

Comparisons of temporal gene expression for genes with high and low correlation were performed by taking the top and bottom 100 genes by Spearman’s correlation (corresponding to correlation > 0.627 and < -0.5032). Estradiol data taken from SRP070657, and DHT data taken from SRP059762. As we consider the expression patterns of many genes, all with different baseline expression, we scaled the expression of every gene by dividing the expression at each time point by the expression at t=0. Plots were generated in ggplot2 using geom_smooth (loess).

Connectivity Map analyses

Connectivity Map L1000 data were downloaded from the NCBI GEO database (GSE70138 and GSE92742). We extracted imputed expression data for TMPRSS2 and ACE2, and for each treatment compound, we performed a non-parametric Mann-Whitney U test between control DMSO-treated samples and drug-treated samples. As Connectivity Map data tests drug treatments across a wide range of concentrations, we performed comparisons using the following dosage groups: (i) any dosage, (ii) below 0.5uM, (iii) 0.2uM to 1uM, (iv) 0.5uM to 2uM, (v) below 1uM. Multiple testing correction was performed using the Bonferroni-Hochberg correction.

To identify broad transcriptional inhibitor of SARS-CoV-2 viral protein interacting partners, we queried the Connectivity Map (clue.io) to identify compounds most likely to down-regulate host proteins required for SARS-CoV-2 pathogenesis as determined via a protein-protein interaction analysis[16]. The Connectivity Map considers up to 150 genes per query. We used three different signatures as input. In the first, we included the 33 out of the 332 host proteins that are directly quantified by the L1000 platform (i.e. “landmark genes”). The second signature included these landmark genes and 33 other well-imputed (self-correlation >0.8) genes. In the third signature, we considered the top 150 genes from the protein-protein interaction map, ranked by fold change. We considered the “summary” Connectivity Score for each compound, which represents the average Connectivity Score achieved in each of the nine core CMap cell lines.

 doi:10.20944/preprints202003.0360.v1

References

1. World Health Organization. (2020). Novel Coronavirus (2019-nCoV): Situation Report 63. Geneva: WHO; 2020. Available from: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200323-sitrep-63-covid-19.pdf?sfvrsn=d97cb6dd_2
2. Guan, W.-J. *et al.* Clinical Characteristics of Coronavirus Disease 2019 in China. *N Engl J Med* (2020). doi:10.1056/NEJMoa2002032
3. Wu, C. *et al.* Risk Factors Associated With Acute Respiratory Distress Syndrome and Death in Patients With Coronavirus Disease 2019 Pneumonia in Wuhan, China. *JAMA Intern Med* (2020). doi:10.1001/jamainternmed.2020.0994
4. Cao, B. *et al.* A Trial of Lopinavir-Ritonavir in Adults Hospitalized with Severe Covid-19. *N Engl J Med* (2020). doi:10.1056/NEJMoa2001282
5. Holshue, M. L. *et al.* First Case of 2019 Novel Coronavirus in the United States. *N Engl J Med* **382,** 929–936 (2020).
6. Xu, Y. *et al.* Characteristics of pediatric SARS-CoV-2 infection and potential evidence for persistent fecal viral shedding. *Nat Med* 1–9 (2020). doi:10.1038/s41591-020-0817-4
7. Mitjà, O. & Clotet, B. Correspondence. *The Lancet Global Health* 1–2 (2020). doi:10.1016/S2214-109X(20)30114-5
8. Grasselli, G., Pesenti, A. & Cecconi, M. Critical Care Utilization for the COVID-19 Outbreak in Lombardy, Italy: Early Experience and Forecast During an Emergency Response. *JAMA* (2020). doi:10.1001/jama.2020.4031
9. Yan, R. *et al.* Structural basis for the recognition of the SARS-CoV-2 by full-length human ACE2. *Science* eabb2762 (2020). doi:10.1107/S0907444910007493
10. Simmons, G. *et al.* Inhibitors of cathepsin L prevent severe acute respiratory syndrome coronavirus entry. *Proc. Natl. Acad. Sci. U.S.A.* **102,** 11876–11881 (2005).
11. Hoffmann, M. *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* 1–19 (2020). doi:10.1016/j.cell.2020.02.052
12. Glowacka, I. *et al.* Evidence that TMPRSS2 Activates the Severe Acute Respiratory Syndrome Coronavirus Spike Protein for Membrane Fusion and Reduces Viral Control by the Humoral Immune Response. *J. Infect. Dis.* **190,** 4122–4134 (2011).
13. Simmons, G., Zmora, P., Gierer, S., Heurich, A. & Pöhlmann, S. Antiviral Research. *Antiviral Research* **100,** 605–614 (2013).
14. Matsuyama, S. *et al.* Efficient activation of the severe acute respiratory syndrome coronavirus spike protein by the transmembrane protease TMPRSS2. *J. Virol.* **84,** 12658–12664 (2010).
15. Ziegler, C. *et al.* SARS-CoV-2 Receptor ACE2 is an Interferon-Stimulated Gene in Human Airway Epithelial Cells and Is Enriched in Specific Cell Subsets Across Tissues. *CellPress Sneak Peek.* Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3555145
16. Gordon, D. E. *et al.* A SARS-CoV-2-Human Protein-Protein Interaction Map Reveals Drug Targets and Potential Drug-Repurposing. *bioRxiv.org* doi:10.1101/2020.03.22.002386
17. Kim, T. S., Heinlein, C., Hackman, R. C. & Nelson, P. S. Phenotypic analysis of mice lacking the Tmprss2-encoded protease. *Mol. Cell. Biol.* **26,** 965–975 (2006).
18. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).
19. Iwata-Yoshikawa, N. *et al.* TMPRSS2 Contributes to Virus Spread and Immunopathology in the Airways of Murine Models after Coronavirus Infection. *J. Virol.* **93,** (2019).

20. Kuba, K. *et al.* A crucial role of angiotensin converting enzyme 2 (ACE2) in SARS coronavirus–induced lung injury. *Nat Med* **11,** 875–879 (2005).
21. Halangk, W. *et al.* Role of cathepsin B in intracellular trypsinogen activation and the onset of acute pancreatitis. *J. Clin. Invest.* **106,** 773–781 (2000).
22. Petermann, I. *et al.* Lysosomal, cytoskeletal, and metabolic alterations in cardiomyopathy of cathepsin L knockout mice. *FASEB j.* **20,** 1266–1268 (2006).
23. Xu, X., Greenland, J. R., Gotts, J. E., Matthay, M. A. & Caughey, G. H. Cathepsin L Helps to Defend Mice from Infection with Influenza A. *PLoS ONE* **11,** e0164501 (2016).
24. Tsui, B., Dow, M., Skola, D. & Carter, H. Extracting allelic read counts from 250,000 human sequencing runs in Sequence Read Archive. *Pac Symp Biocomput* **24,** 196–207 (2019).
25. Mikkonen, L., Pihlajamaa, P., Sahu, B., Zhang, F.-P. & Jänne, O. A. Androgen receptor and androgen-dependent gene expression in lung. *Molecular and Cellular Endocrinology* **317,** 14–24 (2010).
26. Haffner, M. C. *et al.* Androgen-induced TOP2B-mediated double-strand breaks and prostate cancer gene rearrangements. *Nature Genetics* **42,** 668–675 (2010).
27. Subramanian, A. *et al.* A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **171,** 1437–1452.e17 (2017).
28. Lim, N. & Pavlidis, P. Evaluation of Connectivity Map shows limited reproducibility in drug repositioning. *bioRxiv.org*
29. Srivatsan, S. R. *et al.* Massively multiplex chemical transcriptomics at single-cell resolution. *Science* **367,** 45–51 (2020).
30. Chen, K. *et al.* Antiinflammatory effects of bromodomain and extraterminal domain inhibition in cystic fibrosis lung inflammation. *JCI Insight* **1,** L257 (2016).
31. Livingston, E. & Bucher, K. Coronavirus Disease 2019 (COVID-19) in Italy. *JAMA* (2020). doi:10.1001/jama.2020.4344
32. Liu, W. *et al.* Detection of Covid-19 in Children in Early January 2020 in Wuhan, China. *N Engl J Med* (2020). doi:10.1056/NEJMc2003717
33. Lu, X. *et al.* SARS-CoV-2 Infection in Children. *N Engl J Med* (2020). doi:10.1056/NEJMc2005073
34. Cai, H. Correspondence. *The Lancet Respiratory* 1–1 (2020). doi:10.1016/S2213-2600(20)30117-X
35. MD, X. Y. *et al.* Clinical course and outcomes of critically ill patients withSARS-CoV-2 pneumonia in Wuhan, China: a single-centered,retrospective, observational study. 1–7 (2020). doi:10.1016/S2213-2600(20)30079-5
36. Zhang, J.-J. *et al.* Clinical characteristics of 140 patients infected with SARS-CoV-2 in Wuhan, China. *Allergy* **395,** 507 (2020).
37. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348,** 648–660 (2015).
38. Ardini-Poleske, M. E. *et al.* LungMAP: The Molecular Atlas of Lung Development Program. *American Journal of Physiology-Lung Cellular and Molecular Physiology* **313,** L733–L740 (2017).
39. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **102,** 15545–15550 (2005).
40. Baran-Gale, J., Purvis, J. E. & Sethupathy, P. An integrative transcriptomics approach identifies miR-503 as a candidate master regulator of the estrogen response in MCF-7 breast cancer cells. *RNA* **22,** 1592–1603 (2016).
41. Plovanich, M., Weng, Q. Y. & Mostaghimi, A. Low Usefulness of Potassium Monitoring Among Healthy Young Women Taking Spironolactone for Acne. *JAMA Dermatol* **151,** 941 (2015).

42. National Heart, Lung, and Blood Institute Acute Respiratory Distress Syndrome (ARDS) Clinical Trials Network *et al.* Comparison of two fluid-management strategies in acute lung injury. *N Engl J Med* **354,** 2564–2575 (2006).
43. Channappanavar, R. *et al.* Sex-Based Differences in Susceptibility to Severe Acute Respiratory Syndrome Coronavirus Infection. *J. Immunol.* **198,** 4046–4053 (2017).
44. Glowacka, I. *et al.* Differential Downregulation of ACE2 by the Spike Proteins of Severe Acute Respiratory Syndrome Coronavirus and Human Coronavirus NL63. *J. Virol.* **84,** 1198–1205 (2009).