SARS-CoV-2 infection mediates differential expression of human endogenous retroviruses and long interspersed nuclear elements

Jez L. Marston, …, Luis P. Iñiguez, Douglas F. Nixon


Graphical abstract

Find the latest version:

https://jci.me/147170/pdf
SARS-CoV-2 infection mediates differential expression of human endogenous retroviruses and long interspersed nuclear elements

Jez L. Marston¹, Matthew Greenig¹, Manvendra Singh², Matthew L. Bendall¹, Rodrigo R. R. Duarte¹, Cédric Feschotte², Luis P. Iñiguez*¹, Douglas F. Nixon*#¹

¹Division of Infectious Diseases, Weill Cornell Medicine, New York, NY 10021, USA.
²Department of Molecular Biology & Genetics, Cornell University, Ithaca, NY 14853, USA.

# Correspondence to:
Douglas F. Nixon (dnixon@med.cornell.edu)
413 E 69th St. Room 530
New York, NY, USA 10021
+1-646-962-2681

*Shared Senior Authors

The authors have declared that no conflict of interest exists.
Abstract [198 words]

Severe respiratory coronavirus 2 (SARS-CoV-2) promotes an imbalanced host response which underlies the development and severity of COVID-19. Infections with viruses are known to modulate transposable elements (TEs) which can exert downstream effects by modulating host gene expression, innate immune sensing, or activities encoded by their protein products. We investigated the impact of SARS-CoV-2 infection on TE expression using RNA-seq data from cell lines and from primary patient samples. Using a bioinformatic tool, Telescope, we showed that SARS-CoV-2 infection led to up- or down-regulation of TE transcripts, a subset of which differed from cells infected with SARS, MERS, RSV, HPIV3 or IAV. Differential expression of key retroelements specifically identified distinct virus families such as coronaviridae, with unique retroelement expression subdividing viral species. Analysis of ChIP-seq data shows that TEs differentially expressed in SARS-CoV-2 infection are enriched for binding sites for TFs involved in immune responses and for pioneer transcription factors. In COVID-19 patient samples, there was a significant TE overexpression in bronchoalveolar lavage fluid and downregulation in peripheral blood mononuclear cells. Thus, while the host gene transcriptome is altered by infection with SARS-CoV-2, the retrotranscriptome may contain the most distinctive features of the cellular response to SARS-CoV-2 infection.
Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection has caused a worldwide pandemic with many millions of people infected. The development and distribution of SARS-CoV-2 specific vaccines has helped to suppress new infections, but a rising number of variant strains threaten global efforts to achieve sufficient protective immunity. Since the identification of SARS-CoV-2 the search for new or repurposed drugs that treat coronavirus disease 2019 (COVID-19) has been disappointing, with a limited number of drugs currently authorized for emergency use. A better understanding of viral immunopathogenesis could identify new avenues for treatments. Current studies have identified viral replication at the beginning of infection and a latter exuberant host immune response as two major opportunities for clinical intervention\(^1\).

Upon a virus infection, cell host and viral genes act separately and in tandem to promote viral replication as well as initiate a host cell response. Coronaviruses, including SARS-CoV-2, hijack host cellular machinery and gene expression networks in order to enhance their own replication within the cell\(^2, 3\). During this process, viral genetic material and proteins serve as signals to elicit innate and adaptive immune responses\(^4-7\). Although a recent study has recognized the co-option of a TE sequence for differential isoform usage in the expression of the ACE2 receptor in the response to SARS-CoV-2 infection\(^8\), most coronavirus studies to date have ignored the potential contribution of transposable elements (TEs) in disease pathogenesis.

Human endogenous retroviruses (HERVs) and long interspersed nuclear elements type 1 (LINE-1) are TEs that contribute numerous repeat sequences throughout the human genome. Despite comprising approximately 25% of the genome, these repetitive
sequences are often disregarded in genomic and transcriptional studies. However, the non-coding and coding activities of these retroelements impinge on a variety of cellular processes that can directly influence physiological responses, including innate sensing, and disease states(9). Epstein Barr Virus(10) and Human Immunodeficiency Virus(11) are two viruses known to modulate TEs with downstream effects on immunopathogenesis, but how coronaviruses, particularly SARS-CoV-2, change TE expression to influence disease pathology and clinical course is a new area of research(12-14).

Previous research has demonstrated significant variation in the expression of HERVs in response to different viral infections in both human and mouse models(15). LINE-1 elements have also been implicated in the pathogenesis and inhibition of viral infections(16). Therefore, we hypothesized that differentially expressed TEs in SARS-CoV-2 infection could reveal new mechanisms underlying the pathogenesis of COVID-19 and direct further research into new therapeutic opportunities. Here, we analyzed publicly available RNA sequencing (RNA-seq) data of cell lines infected with the SARS-CoV-2, severe acute respiratory syndrome coronavirus (SARS-CoV or SARS), middle east respiratory syndrome coronavirus (MERS-CoV or MERS), Influenza A Virus (IAV), Respiratory Syncytial Virus (RSV) and Human Parainfluenza Virus Type 3 (HPIV3) to determine virus-specific signatures in the retrotranscriptome. We further analyzed RNA-seq data obtained from clinical samples of patients infected with SARS-CoV-2. We found that differential expression of key retroelements could identify a virus family such as coronaviridae, with unique retroelement expression subdividing viral species. Thus, the
retrotranscriptome may contain the most distinctive features of host response to SARS-CoV-2 infection and provide new insights into pathogenesis and treatment of COVID-19.
Results

Differential expression of transposable elements after SARS-CoV-2 infection

In order to elucidate the effects of SARS-CoV-2 infection on the human retrotranscriptome we quantified the expression of TEs from publicly available RNA-seq data of SARS-CoV-2 infected pulmonary cell lines (17, 18) and primary COVID-19 patient-derived samples (19) using the specialized bioinformatic tool ‘Telescope’ (20) (Table 1, Table S1). Telescope applies an Expectation-Maximization algorithm to reassign multiple mapped reads to a specific location allowing a genomic locus-specific resolution of TE expression. We assessed differential expression (DE) of retroelements, as well as genes, and compared retrotranscriptomic changes from SARS-CoV-2 infection to those induced by related pathogenic human coronaviruses, SARS and MERS, and by other common respiratory viruses, IAV, RSV, and HPIV3. As described in Blanco-Melo et al. (17), A549 cells were found to be relatively non-permissive to SARS-CoV-2 infection in contrast to Calu-3 cells given their lower expression of the SARS-CoV-2 host receptor ACE2. We therefore also compared the retrotranscriptomes of A549 cells transduced with a vector expressing human ACE2 receptor (A549-ACE2) at different SARS-CoV-2 multiplicities of infection (MOI) in order to determine the effects of viral burden on the expression of retroelements. We also compared the retrotranscriptomes of A549 cells infected at the higher MOI with those pre-treated with the kinase inhibitor ruxolitinib to determine the effects of IFN-1 signaling and JAK-STAT pathway interference on retroelement expression. Blanco-Melo et al (2020) showed that inhibition of this pathway reduced the
expression of interferon stimulated genes (ISGs) and enhanced viral replication, however this intervention did not inhibit the release of inflammatory cytokines (17).

We first performed a principal component analysis (PCA) on the transcriptomic and retrotranscriptomic data from Calu-3, A549 and A549-ACE2 cells, and found that the infected samples clustered separately depending on the type of virus infection, even between closely related coronaviruses (Fig. S1). This result demonstrates that infected cells not only undergo virus-specific changes in gene expression, but also in their expression of HERVs and LINE-1 retroelements.

In Calu-3 cells, the three coronavirus infections (SARS-CoV-2, SARS and MERS) each elicited unique profiles of retroelement expression (Fig. 1A-D, Table S2). For SARS-CoV-2 infection there were 92 upregulated retroelements (52 HERVs and 40 LINE-1 elements), of which 14 HERVs and 7 LINE-1 elements were identified to be intergenic.

Since many retroelements intersect human genes, the upregulation of intergenic retroelements indicates the presence of transcripts uniquely derived from retroelement sequences in response to infection. Of note, ERV316A3_6p22.3c, an HERV that was upregulated only with SARS-CoV-2 infection, intersects HCP5, a long non-coding RNA (IncRNA) previously reported to be driven by the HERV promoter (21). Additionally, a set of 72 retroelements were upregulated across all three coronavirus infections including 40 HERVs and 32 LINE-1 elements, and of these, 13 HERVs and 4 LINE-1 elements were intergenic (Fig. 1D).

We performed a TE family enrichment of all elements differentially expressed after each coronavirus infections. The results showed a shared enrichment of retroelements from the HERVE and HARLEQUIN families while SARS-CoV-2 specifically enriched for
differential expression of HERVH and HERV3 loci (adjusted p < 0.05) (Fig. S2A-B).

Comparing the differential expression of these enriched HERV families showed that each virus induced unique changes in expression at these loci, including those that were upregulated (HERVH_4q31.3c) or downregulated (HERVH_11q24.1c) across all three infections, as well as those specific for SARS-CoV-2 including HERVH_15q26.3b, HARLEQUIN_7q33a, HERV3_7q33, HERVH_13q33.3, and HERVH_4p15.33b, among others (Figure S2B).

In A549 cells, SARS-CoV-2, RSV, HPIV3 and IAV all induced differential expression of individual TEs. As with Calu-3 cells, each viral infection in A549 cells was associated with its own virus-specific retrotranscriptomic profile. Interestingly, the greatest number of DE TE loci occurred with SARS-CoV-2 infection, followed by RSV then HPIV3, despite low permissiveness to infection. Intriguingly, IAV infection only induced limited changes in TE expression (Figure 1E-I, Table S3). Two elements had differential expression for all four viral infections - MER4_22q12.3 and L1FLnl_7q21.2o - which intersect with the human genes APOL6 and SAMD9, respectively. MER4_22q12.3 contributes to the 3’ UTR of APOL6 and therefore may provide a polyA signal to the end of this ISG. The L1FLnl_7q21.2o sequence is antisense to SAMD9 but contributes a large portion of the gene’s intron and second exon. Both genes are IFN-γ responsive which may explain the expression of intersecting retroelements, however the contributions of these retroelement sequences to these ISGs remains to be studied (22, 23). Ten retroelements were differentially expressed in SARS-CoV-2, RSV or HPIV3 infections, but not with IAV (Fig. 1I). These include three elements that belong to the ERV316A3 family located on chromosomes 2, 6, and 12, including the HCP5-intersecting locus. The
A unique response to SARS-CoV-2 infection includes the overexpression of 91 elements of which 51 were HERVs and forty were LINE-1 elements.

In A549-ACE2 cells infected with SARS-CoV-2 there was a substantial increase in differentially expressed TEs in comparison to non-transduced A549 cells (Figure 1J-M, Table S4). Of note, a larger number of TE loci were upregulated than downregulated, as observed in other cell lines. A comparison of A549-ACE2 cells infected with SARS-CoV-2 at two different MOI showed that more than 100 TEs were differentially expressed at either MOI, but there was a greater differential expression at the higher MOI. One hundred and forty-five retroelements were upregulated at an MOI of 2 not seen at the lower MOI of 0.2, including 17 intergenic HERVs and 8 intergenic LINE-1 elements. HCP5 was shown to be upregulated at the higher and not the lower MOI, suggesting a threshold of viral burden to transactivate its expression. Administration of the JAK-STAT pathway inhibitor ruxolitinib (Rux) prior to SARS-CoV-2 infection also altered the TE expression response, but the majority of differentially expressed retroelements were shared across the different MOI and treatment conditions. However, ruxolitinib treatment inhibited the upregulated expression of 21 retroelements, including 6 HERVs and 15 LINE-1 elements. Of the 6 HERVs, 4 were intergenic sequences with no overlapping human coding genes and an annotated HERV-derived gene ERV3-1, overlapping with the ZNF117 gene and HERV4_4q22.1. The changes in the expression response of these intragenic HERV transcripts suggests independent transcriptional regulation of these loci by JAK-STAT signaling pathways.

While each model system showed distinct profiles, there were shared elements in all three in vitro models of untreated SARS-CoV-2 infection (Calu-3, A549 and A549-
ACE2 at both MOIs) with upregulation of 18 common retroelements including 9 HERVs and 9 LINE-1 elements, of which 3 HERVs were intergenic, including HARLEQUIN_7q33a, HERV3_7q33 and HML6_6p22.2, suggesting locus-specific transactivation of these HERVs upon SARS-CoV-2 infection. Of note, both ERV316A3_6p21.33c (HCP5) and MER4_22q12.3 (3’ UTR of APOL6) were upregulated across all three cell models of SARS-CoV-2 infection (Figure S3, Table S5).

Associations between the SARS-CoV-2 induced retrotranscriptome and innate immune genes

Given the significant dysregulation of key cytokines such as IL6 and TNFα in COVID-19 patients, we analyzed the RNA-seq data from these in vitro infection models and compared to samples from COVID-19 patients for similar immune gene expression signatures. SARS-CoV-2 significantly upregulated the expression of downstream NF-κB-regulated innate immune genes such as IL6, IL1B, TNF in Calu-3 cells (24). However, upregulation of these IL1B transcripts was not seen in SARS or MERS infection of Calu3 cells (Fig. S4). We also compared 3 host immune genes with intersecting HERV loci differentially expressed in SARS-CoV-2 infection, APOL6, HCP5, and SP140L (Fig. S4).

We show that these intersected genes were also differentially upregulated in SARS-CoV-2 infection alongside the identified retroelement loci, MER4_22q12.3, ERV316A3_6p21.33c and ERV316A3_2q37.1a, respectively.

We then explored associations between the SARS-CoV-2 induced retrotranscriptome and key innate immune gene expression changes in host cells. We applied the GREAT Analysis tool, which assigns biological meaning to a set of non-
coding, or TE, regions by analyzing nearby gene annotations, to the differentially expressed TE sequences induced by SARS-CoV-2, SARS and MERS infection in Calu-3 cells (25). From this analysis, GO terms were derived from enriched host genes. There was a marked enrichment for GO terms pertaining to type I interferon signaling pathway, defense response to virus, innate immune response and IFN-γ-mediated signaling pathway terms (Table S5). The differentially expressed genomic regions were associated with numerous immune genes including OAS2, OAS3, IFNAR1, IL10RB, SP100, HLA-B, HLA-G, HLA-V, FCGR2A, IL6, APOL5, APOL6, JAK2, MX1 and MX2 (Table S6). Despite this enrichment in SARS-CoV-2 infection, no GO terms were enriched from differentially expressed TEs from SARS or MERS infections. The set of differentially expressed retroelement loci from the other cell types did not show enrichment for specific GO sets through the GREAT analysis. This observation may be due to differences in permissibility to SARS-CoV-2 infection as noted in Blanco-Melo et. al, 2020 (17).

Immune response genes and specific retroelements are upregulated in SARS-CoV-2 infection

To explore the specific gene and retroelement networks that underlie the host response to SARS-CoV-2 infection, we applied hierarchical clustering to identify sets of genes ("clusters") with significantly correlated expression patterns between the infections of Calu-3, A549 and A549-ACE2 cells. We filtered all genes based on statistically significant differential expression patterns identified by the likelihood ratio test for infection (adjusted p < 0.05), which produced 31 clusters of genes and co-expressed retroelements in Calu-
3 cells, 13 clusters in A549 cells and 18 clusters in A549-ACE2 cells (Fig. 2 and Figs. S5-7).

Importantly, each clustering analysis highlighted a cluster of genes and retroelements that were induced by SARS-CoV-2 infection and enriched for GO terms including “defense response to virus,” “response to interferon gamma,” and “cellular response to type I interferon.” Additionally, these clusters also enriched for KEGG pathways induced by virus infection including IAV, HCV, EBV, HSV-1 and measles (Fig. 2A-C). In A549-ACE2 cells, treatment with ruxolitinib decreased expression of retroelements associated with human genes involved in the response to virus cluster 12. This suggested coordinated expression of these retroelements in the signaling pathways involved in this immune response (Table S7). From the Calu-3 cluster, 4 intergenic HERVs were correlated with antiviral gene expression, ERV316A3_12q24.13 (between to OAS2 and OAS3 genes), ERV316A3_4p15.2c, HERV10F_14q11.2c, and HML1_10q22.1 (Fig. 2D-F). In A549 cells, 2 intergenic HERVs correlated with the antiviral gene cluster expression, HERV3_11q13.3, annotated as LINCO2701, and HERVL_9p24.1b. No intergenic HERVs were associated with the antiviral response in A549-ACE2 cells (Fig. 2G-I).

Transcription factor interactions with differentially expressed TE loci

To further investigate the independent expression of TEs in response to viral infection, we first analyzed differentially expressed TEs seen in SARS-CoV-2 infection of Calu-3 cells for binding of transcription factors based on ChIP-seq data produced as part of the ENCODE project(26). We observed an enrichment of POLII peaks within the TE
sequences themselves, suggesting that the TEs are expressed independently of overlapping host genes. Transcription factors identified in the response to SARS-CoV-2 infection such as JUND, FOSL2, and FOXA2, and chromatin double strand break and remodeling factors, such as CTCF, YY1 and RAD21, were enriched for binding to the TE loci identified from our Telescope analysis (Figure 3A-D). We focused on binding enrichment for pioneer transcription factors that bind closed chromatin and induce chromatin opening, such as FOXA and GATA3(27, 28). Of the DE retroelements, the top 20 retroelements with enrichment for pioneer factor binding sites in A549 cells included HERV3_7q33, HERVH_13q33.3 and HML6_6p22.2 which were significantly upregulated across SARS-CoV-2 infections. (Table S8).

We next used ChIP-seq data for the histone acetylation mark H3K27Ac, which correlates with transcriptional activity(29), to profile the regions surrounding the DE TE loci during SARS-CoV-2 infection of A549 cells(30). We observed an increase accumulation of this histone mark over time in the region immediately upstream of the differentially expressed TEs, suggesting that chromatin remodeling of these TEs to an active state after 24 h of SARS-CoV-2 infection (Figure 3E). This pattern is observed for a diverse set of TEs and genomic contexts including an ERV316A3 element that contributes to HCP5, an HERVE element downstream of the B3GNT7 gene, a MER4 element within the 3’UTR of APOL6 and four intergenic HERV sequences, HERVH_13q33.3, HARLEQUIN_7q33a, HERV3_7q33 and HML6_6p22.2 (Figure S8A-D).

In addition, we analyzed the compendium of publicly available data of 1000 ChIP-seq derived transcription factor (TF) binding sites provided by the ReMap2020(31). By
comparing the annotated transcription factor binding sites (TFBS) within the genomic regions of the SARS-CoV-2 induced retroelements against the full set of TFBSs with all retroelement regions in the genome, we noted an enrichment of binding sites for TFs involved in immune responses (REL, IRF5, and BCL3), chromatin structural regulation (MPHOSPH8, CHAMP1, TRIM28) and transcriptional activation (JUND and TBP). This comparison further suggests that these retroelements are subject to chromatin remodeling upon infection (Fig. S9).

In order to investigate the viral specificity of this epigenetic response we then compared H3K27ac data over the time course of H1N1 influenza virus infection of Human Bronchial Tracheal Epithelium (HBTE) cells, a primary cell model for pulmonary viral infection(32). Similar to our findings for SARS-CoV-2 infection in A549 cells, we found an increase of this histone mark over time in the region upstream of the same set of differentially expressed TEs (Figure 3E). Lastly, the HERVE_2q37.1 integrant downstream of the B3GNT7 gene is also acetylated after influenza infection, and harbors the regulatory sequence of its neighboring human gene. A subset of DE TE loci undergo similar epigenetic changes after influenza and SARS-CoV-2 infection, possibly leading to their overexpression.

SARS-CoV-2 induced retrotranscriptome changes in primary clinical samples

In order to explore whether these findings were replicated in clinical samples, we analyzed publicly available bulk RNA-seq data of bronchoalveolar lavage fluid (BALF) and peripheral blood mononuclear cells (PBMCs) from patients hospitalized with severe COVID-19(19) compared with samples from healthy donor controls. The results showed
a profound dysregulation in the expression of retroelements (Fig. 4A-D), with significant overexpression in BALF and downregulation in PBMC. Conducting a principal component analysis based upon these retrotranscriptomic profiles, the infected samples clustered separately from the healthy clinical samples, with the infected BALF samples showing more profound changes than those of PBMC (Fig. S10, Table S9). Comparing across lung tissue-derived samples (Calu-3, A549, A549-ACE2 and BALF), one LINE-1 transcript was differentially expressed across all SARS-CoV-2 infections, L1FLnL_16p11.2i (Fig. 4E), however, this observed change in expression may be due to differential expression of the ZNF267 gene within which this LINE-1 element is contained.

Given recent findings of differential expression of a pathogenic HERVW in the leukocytes of COVID-19 patients(12), we sought to determine whether the PBMC may show differential expression of HERVW transcripts. In PBMC, we saw a marked upregulation of HERVW_15q21.2 which intersects the GLDN gene. However, no ORFs encoding an envelope protein was determined from this locus. Interestingly, BALF, which may also contain numerous immune cells, showed a marked upregulation of HERVW_7q21.2, also known as ERVW-1, which encodes an envelope protein, syncytin-1, known to mediate the formation of syncytia, specifically in placental syncytiotrophoblasts(33, 34). Another HERW locus, HERVW_4q21.22, which lies intergenic between ENOPH1 and TMEM150C, was upregulated in BALF and can produce an annotated transcript, XM_011532442.2, which is annotated as a “protein-coding”, “syncytin-1-like” transcript, LOC105377310. The transcript contains an open reading frame for a 245-residue envelope protein with 85.37% sequence homology to syncytin-1 by BLAST search. Similarly, HERVFRD_6p24.2 which encodes syncytin-2, an
envelope protein from the HERVFRD family also involved in placental syncytiotrophoblast fusion (35), was upregulated in the BALF samples of COVID-19 patients. Thus, our analysis suggests that several HERV protein products may be overexpressed in BALF of COVID-19 patients.
Discussion

In this study, we describe the impact of SARS-CoV-2 infection on the transcription of locus-specific endogenous retrotransposons, specifically HERV and LINE-1 elements. This level of retrotranscriptomic precision provides insights into the potential genomic effects of exogenous viral and host immune factors on the activation of endogenous retrotransposon sequences. Given the extensive literature on the cooption of endogenous retroviruses as immune-responsive cis-regulatory elements (36-38) this study provides a quantitative analysis of retrotransposon expression patterns from publicly available RNA-seq data in response to viral infections, with a focus on the specific TE responses to SARS-CoV-2 infection.

We quantified and analyzed the retrotranscriptome of SARS-CoV-2 infection in *in vitro* cell lines, in comparison to SARS and MERS and to other common respiratory viruses IAV, RSV and HPIV3. SARS-CoV-2 induced the expression of a unique retrotranscriptomic signature in comparison to related coronavirus infections in Calu-3 cells, and to other common respiratory virus infections in A549 cells. The retrotranscriptome of SARS-CoV-2 differed between each cell type suggesting that retrotranscriptomic activation in response to virus is both virus-specific and cell type-specific. Furthermore, there was marked upregulation of TEs in BALF, but not from PBMCs from COVID-19 donors.

In SARS-CoV-2 infected A549-ACE2 cells, a higher MOI induced up- and downregulation of a larger number of retroelements, suggesting that the induction of certain elements may be dependent on the level of viral burden. Pre-treatment of A549-
ACE2 cells with ruxolitinib also showed many similarities to untreated cells, suggesting that the retrotranscriptomic response to SARS-CoV-2 in the A549-ACE2 cells was largely reproducible and consistent and that the expression of a small number of retroelements was altered by pretreatment with the kinase inhibitor. However, many of these elements intersected known interferon and immune response genes, which may account for the changes in expression of these retroelements. The finding that ruxolitinib effectively reduced the expression of ISGs in vitro is significant in the context of SARS-CoV-2 infection and this pharmacological intervention has being tested in the RUXCOVID clinical trial which did not meet the primary endpoint of reducing the number of hospitalized COVID-19 patients who experienced severe complications (39).

For this study, although the expression of retroelements in response to SARS-CoV-2 infection was largely unchanged by pretreatment with ruxolitinib, there is an interesting similarity in unchanged cytokine expression found in the Blanco-Melo study (17). This finding may suggest that SARS-CoV-2 induced retroelements correspond more with inflammatory cytokine expression dynamics than with the interferon stimulated response. Perhaps, the expression of these elements may themselves be cytokine-like mediators of inflammation (40).

In Calu-3 cells we showed that SARS-CoV-2 induced an elevated expression of HERVs from the HERVH and HERV3 families (not enriched in SARS or MERS infections), and shared enrichment of HARLEQUIN and HERVE families across virus types. This result suggests a common induction of TE families after coronavirus infections and specifically for SARS-CoV-2. Additionally, using the GREAT Analysis tool, we showed that the set of retroelements that were differentially expressed in
response to SARS-CoV-2 infection of Calu-3 cells were enriched in genomic regions proximal to human genes involved in the innate immune response, response to virus and the response to interferon gamma pathways. This finding suggests that these elements may play roles in the expression or modulation of these immune genes (17, 36) or may themselves be transcribed alongside these immune response genes. This genomic enrichment was unique to the differentially expressed retroelements induced by SARS-CoV-2 in Calu-3 cells and not in other viral infections or in the A549 or A549-ACE2 infections with SARS-CoV-2, which may be due to the differences in A549 permissibility to SARS-CoV-2 infection (17).

A set of upregulated retroelements were shared across the cell-types infected with SARS-CoV-2 which implies that these elements may play a role in the infection cycle or host response to SARS-CoV-2 infection. Many of these elements were also intergenic elements and may be activated independently from host genes by SARS-CoV-2 infection. Of particular note, the HARLEQUIN_7q33a and HERV3_7q33 transcripts, which may be expressed as individual or mosaic transcripts, were upregulated across all infections. The same HERV3 locus has previously been described as a marker of response to immunotherapy in various cancers (41) and lies downstream of the CYREN gene, of which an alternatively spliced transcript encodes the modulator of retrovirus infection (MRI-1) micropeptide (42). Additionally, an intergenic HERVK transcript, HML6_6p22.2, was also upregulated across the SARS-CoV-2 infected cells. The HERVK family has been implicated in various disease pathologies (43-45). Further investigation will seek to characterize the expression, regulation and coding potential of these elements to determine their contributions to
physiology and disease. Finally, this analysis points a spotlight on two human genes

$HCP5$ and $APOL6$ whose transcripts contain HERV-derived sequences and are
upregulated in SARS-CoV-2 infection. Both genes have been shown to be involved in
the immune response, but further characterization is warranted in connection to SARS-
CoV-2 infection and the contribution of their intrinsic HERV sequence components(46).

By applying a likelihood ratio test (LRT) to determine differentially expressed
genomes for downstream co-expression clustering and analysis, we showed that
differentially expressed retroelements co-clustered with many gene networks in
response to the various viral infections in each cell type tested. Notably, we identified
the set of retroelement transcripts that were co-expressed with clusters that were
enriched for genes involved in the defense response to virus, response to type I
interferon and response to interferon gamma GO terms and KEGG pathways related to
various viral infections. This adds to the hypothesis that certain retroelements may be
co-activated with or play roles in the immune response to virus.

One common caveat of the current analysis - and of retrotranscriptomic analysis
in general(47) - is that many retroelement sequences are nearby or overlap with
previously described protein-coding or noncoding genes. Thus, expression of TE loci
may not be mechanistically related to retroelement sequences but in fact derive from
transcriptional regulation of nearby genes. However, we posit that observed expression
of TE loci can be and often is a property of the TE locus and should be considered
together with analysis of other genes. First, we have shown that repetitive TE loci are a
rich source of cis-regulatory elements(4). Second, TE sequences themselves encode
functional elements, including peptides and RNAs. For example, HCP5 is an antisense
retroviral transcript that has been coopted by the innate immune system through an
HLA class I promoter (48). Many of the differentially expressed retroelements were also
intergenic, suggesting that changes in transcription were independent of any intersected
human genes.

We also addressed whether these retroelements were being independently
transcribed in response to viral infection by analyzing the transcriptional and epigenetic
regulation of these genomic sequences in response to virus infection. From an analysis
of transcription factor binding site enrichment for the set of retroelements differentially
expressed in SARS-CoV-2 infection of Calu-3 cells, we showed that a portion of these
elements had extensive ChIP-seq peaks from transcription initiation factors, suggesting
the potential for independent transcription. Additionally, this enrichment was seen for
the same retroelements induced in available data from A549 cells. We then compared
histone acetylation marks (H3K27Ac) to analyze chromatin alterations around the set of
differentially expressed retroelements in publicly available ChIP-seq data of SARS-CoV-
2 infection of A549 cells and, to assess the potential for broad response to other
viruses, H1N1 infection of primary HBTE tissue. We showed that after various time
points after infection, the histone acetylation changes around these differentially
expressed retroelements in both SARS-CoV-2 infection of A549 cells and H1N1
infection of HBTE, with a notably increase in the size of the H3K27Ac peaks after 18-24
hpi. We also demonstrated that the genomic sequences of the differentially expressed
retroelements in SARS-CoV-2 were enriched for the NF-kB subunits RELA and RELB,
implicating this immune response pathway in modulating retroelement expression.
By analyzing the retrotranscriptomes of primary BALF and PBMC samples from hospitalized COVID-19 patients, we showed that profound changes in the retrotranscriptome correlate with disease. In the BALF there was a marked upregulation of retroelement expression, similar to that seen in cell lines, and some of these elements are shared across the in vitro lung adenocarcinoma cell infections. In contrast to the BALF samples, there was a marked downregulation of retroelement expression in the PBMC from COVID-19 patients. This result was striking and may be explained by differences in permissibility to viral infection of blood cells. In the case of low SARS-CoV-2 infection of blood cells, this differential expression may instead be induced by either the inflammatory or cytokine enriched environment resulting from infection of other permissible tissues. This hypothesis should be tested with further studies of blood retrotranscriptomes in response to inflammation or isolated viral infection. While the BALF shared some upregulated retroelements with the in vitro models, the PBMC showed less similarity, suggesting tissue specificity in retroelement induction.

From our analysis of clinical samples of COVID-19, we also show marked upregulation of HERV-derived envelope transcripts encoding syncytin-1 and syncytin-2, as well as another HERVW transcript with an ORF encoding an envelope protein in BALF. In light of findings of a HERVW envelope protein detected in leukocytes of COVID-19 patients(12), the presence of syncytium in infected lung tissue(49), and the fusogenic properties of HERV-derived synctin-1 and synctin-2(33-35), we present data that might support the presence of HERVW envelope transcription in response to infection. These HERV transcripts may play roles in the development of syncytium and the pathophysiology of the inflammatory milieu in COVID-19. Although these proteins
are immunogenic, many retroviral envelope proteins also have been noted to have immunosuppressive functions (50, 51). While the immunosuppressive function of syncytin-1 is still debated, syncytin-2 has been demonstrated to dampen cytokine production (52). Therefore, the presence of these retroviral transcripts and their protein products may also contribute to the delayed interferon response noted in SARS-CoV-2 infection and the inflammatory dynamics of COVID-19 (17). Only one HERVW transcript was shown to be upregulated in PBMC from COVID-19 patients when compared to healthy donors, however, the transcript did not contain an ORF for an envelope protein. This finding may be impacted by the heterogeneity of the PBMC samples and the lower permissibility to SARS-CoV-2 in peripheral blood cells. A major limitation of these findings, however, is that the BALF samples from healthy donors were also from a different study to those derived from COVID-19 patients and may be impacted by batch effects.

In summary, differential expression of transposable elements after viral infections could provide insights into disease pathogenesis and unique viral signatures in viral diagnostics. As the biology of locus specific TEs develops, the interplay of TEs, genes and viral effects will add significant knowledge to the biology of viral infections.
Methods

Samples

In order to compare the expression of host gene and retroelement transcripts from different viral infection models, publicly available data from *in vitro* infection models of SARS-CoV-2 infection in Calu-3 cells described by Blanco-Melo et al. (17) was retrieved from NCBI-SRA with accession numbers: SRR11517744, SRR11517745, SRR11517746, SRR11517747, SRR11517748, SRR11517749. Infection models of A549 cells with SARS-CoV-2, IAV, RSV and HPIV3 were retrieved with accession numbers: SRR11517674, SRR11517675, SRR11517676, SRR11517677, SRR11517678, SRR11517679, SRR11517680, SRR11517681, SRR11517682, SRR11517683, SRR11517684, SRR11517685, SRR11517686, SRR11517687, SRR11517688, SRR11517689, SRR11517690, SRR11517691, SRR11517692, SRR11517693, SRR11517694, SRR11517695, SRR11517696, SRR11517697, SRR11517698, SRR11517699, SRR11517700, SRR11517701, SRR11517702, SRR11517703, SRR11517704, SRR11517705, SRR11517706, SRR11517707, SRR11517708, SRR11517709, SRR11517710, SRR11517711, SRR11517712, SRR11517713, SRR11517714, SRR11517715, SRR11517716, SRR11517717, SRR11517718, SRR11517719, SRR11517720, SRR11517721, SRR11517722, SRR11517723, SRR11517724, SRR11517725, SRR11517726, SRR11517727, SRR11517728, SRR11517729, SRR11517730, SRR11517731, SRR11517732, SRR11517733, SRR11517734, SRR11517735, SRR11517736, SRR11517737, SRR11517738, SRR11517739, SRR11517740, SRR11517741, SRR11517742, SRR11517743, SRR11517744, SRR11517745, SRR11517746, SRR11517747, SRR11517748, SRR11517749. Infection models of SARS-CoV-2 infected A549-ACE2 cells and those pretreated and untreated with ruxolitinib at MOI of 2, were accessed with accession numbers: SRR11573892, SRR11573893, SRR11573894, SRR11573895, SRR11573896, SRR11573897, SRR11573898, SRR11573899, SRR11573900, SRR11573901, SRR11573902, SRR11573903, SRR11573904, SRR11573905, SRR11573906, SRR11573907, SRR11573908, SRR11573909, SRR11573910, SRR11573911, SRR11573912, SRR11573913, SRR11573914, SRR11573915, SRR11573916, SRR11573917, SRR11573918, SRR11573919, SRR11573920, SRR11573921, SRR11573922, SRR11573923, SRR11573924, SRR11573925, SRR11573926, SRR11573927. Samples were run in quadruplicate, so raw fastq files for each sample were concatenated before alignment. RNA-seq data of Calu-3 infection models of SARS and MERS as
described in Yeung et al.(18) was retrieved from NCBI-SRA with accession numbers: SRR1942929, SRR1942934, SRR1942956, SRR1942957, SRR1942960, SRR1942961, SRR1942989, SRR1942990. Finally, RNA-seq data from clinical BALF samples of 2 COVID-19 patients and PBMCs of 3 COVID-19 patients were downloaded from the National Genomics Data Center-Genome Sequence Archive (NGDC-GSA) with accession numbers: CRR119890, CRR119891, CRR119892, CRR119893, CRR119894, CRR119895, CRR119896, CRR119897, CRR125445, CRR125446. RNA-seq data from BALF samples of healthy individuals(53) were downloaded from NCBI-SRA with accession numbers: SRR10571724, SRR10571730, SRR10571732, as compared in Xiong et. al(19).

Transcriptome and Retrotranscriptome quantification
RNA-seq datasets were downloaded and mapped to Gencode V.33 transcriptome with Salmon for gene quantification. The retrotranscriptome expression was determined using Telescope. Briefly, reads were mapped with Bowtie2 allowing up to 100 multi-mapping positions and adjusting the parameters for local alignments (-very-sensitive-local, -k 100 –score-min L,0,1.6). Then, the mapping results were used as input for Telescope using a publicly available annotation of TEs, including HERV and LINE-1 sequences, (https://github.com/mlbendall/telescope_annotation_db/tree/master/builds/retro.hg38.v1) for read reassignment. High-confidence transcript reads assigned to TE sequences in the annotation were used for further analysis.

Filtering genes for clustering
To identify genes exhibiting significant expression variation between experimental conditions, the likelihood ratio test (LRT) from DESeq2 version 1.24.0 was applied separately to the set of retroelement and human transcript counts. For the transcriptomic data derived from Calu-3 cells, the likelihood of a full model containing coefficients for the four different infection conditions was compared to a reduced model containing only an intercept term for each gene. For the A549 cell data, a full model containing coefficients for the five conditions and sequencing batch was compared to a reduced model containing only the sequencing batch coefficient. Finally, this analysis was performed on A549-ACE2 cell data. In all analyses, genes with adjusted p-values < 0.05 were selected for downstream analyses. Human genes and retroelements that were deemed significant by this standard were combined into a single transcript count matrix for downstream analysis. Transcript counts for these genes were stabilized using DESeq2’s variance stabilizing transformation and scaled to zero mean and unit variance within each gene.

Clustering

The set of genes identified as statistically significant by the LRT were clustered using average-linkage agglomerative hierarchical clustering with Pearson correlation distance (1 – Pearson’s correlation) as a dissimilarity metric. A dynamic tree cut(54) was subsequently applied to produce clusters of genes from the dendrogram, using a minimum cluster size of 200 genes for the A549 data and 300 genes for the Calu-3 data.
Cluster expression patterns were visualized using the R package ggplot2(55), plotting each gene’s mean scaled expression value within each experimental condition. Cluster eigengenes were solved as the first principal component of each cluster’s scaled expression matrix. Cluster hub genes and hub retroelements were identified by calculating all pairwise Pearson correlations between genes in the cluster. The top n hub genes for each cluster were identified as the n human genes with the greatest mean correlation with all other genes - both human genes and retroelements - in the cluster. The hub retroelement in each cluster was identified as the retroelement with the greatest mean correlation with all other genes - both human genes and retroelements - in the cluster.

Overrepresentation analysis
Gene clusters were tested for overrepresentation of pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) terms using the hypergeometric test implementation from the R package clusterprofiler(56) with p-value cutoff of 0.05 and a q-value cutoff of 0.20. For each cluster selected for overrepresentation analysis, p-values for the top 5 KEGG pathways and GO terms - ranked by adjusted p-value - were visualized in a heatmap using the R package pheatmap. Terms and pathways that were also significantly overrepresented in other clusters (adjusted p < 0.05) were colored on the heatmap accordingly.

GREAT Analysis
Differentially expressed retroelements from each of the infections and cell types were input as bed files into the GREAT analysis webtool(25). Only the set of differentially expressed retroelements from SARS-CoV-2 infection of Calu-3 cells showed proximity to genomic regions enriched for genes involved in GO terms associated with immune function.

ReMap2020 Transcription Factor Binding Enrichment Analysis

The differentially expressed retroelements were then transformed to genomic ranges objects (GRanges) using the Granges R package. Additionally, all retroelement loci in the Telescope annotation were transformed into GRanges objects. Using the ReMap2020 dataset and R package, the differentially expressed retroelement loci were compared against the ReMap2020 annotated transcription factor bindings sites within all Telescope annotated retroelement loci for enrichment of specific transcription factors and plotted as enrichment dotplots with the ReMap2020 R package.

ENCODEx ChIP-seq Analysis (Manu)

ChIP-seq datasets from ENCODE consortium, SARS-CoV-2 (GSE167528), and H1N1 infection (GSE113702) in sra format were downloaded and converted to fastq format using sratools function fastq-dump --split-3. Fastq reads were mapped against the hg19 reference genome with the modified bowtie2 parameters: --very-sensitive-local. All unmapped reads, the mapped one with MAPQ < 10 and PCR duplicates were removed using Picard and samtools. MACS2 called all the ChIP-seq peaks with the parameters –nomodel -q 0.01 -B. Blacklisted regions were excluded from called peaks
(https://www.encodeproject.org/annotations/ENCSR636HFF/). To generate a set of unique peaks, we merged the peaks using the mergeBed function from bedtools, where the distance between peaks was less than 50 base pairs. We then intersected these peak sets with the differentially expressed TEs from hg19 repeat-masked coordinates using bedtools intersectBed with a 50% overlap. To calculate the enrichment over the given repeat elements, we first extended 5 Kb upstream and 5 Kb downstream coordinates from the left boundary (TSS) of respective elements in a strand-specific manner. These 10 Kb windows were further divided into 100 bps bins and tags (bedGraph output from MACS2) were counted in each bin. Tag counts in each bin were normalized by the total number of tags per million in given samples and presented as CPM per 100 bps. We averaged CPM values for each bin between replicates before plotting the figures.

**Code availability**

All the codes used in the analysis and to generate figures will be available upon acceptance.

**Study approval**

There is no human subjects research in this study.
Acknowledgments

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. J.L.M. was supported in part by a Medical Scientist Training Program grant to the Weill Cornell–Rockefeller–Sloan Kettering Tri-Institutional MD-PhD Program (T32GM007739). R.R.R.D. received funding from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (BEX1279/13-0). This work was supported in part by a Cornell Presidential Postdoctoral Fellowship to MS and funds from the National Institutes of Health to CF (U01HG009391; R35GM122550) and a seed grant from Cornell Office of Academic Integration to CF and DFN.
Author contributions

J.L.M., M.G., L.P.I. and D.F.N. designed the study and wrote the manuscript. M.G., M.S., M.L.B., R.R.R.D. and C.F. helped with study design, performed the data analysis, and wrote the manuscript. J.L.M., M.G., M.L.B., L.P.I. and D.F.N. conceived the research. All authors wrote and approved the manuscript.


22. Zhaorigetu S, Yang Z, Toma I, McCaffrey TA, and Hu CA. Apolipoprotein L6, induced in atherosclerotic lesions, promotes apoptosis and blocks Beclin 1-


Figure 1.

Differential retrotranscriptome profiling of in vitro viral infection of lung adenocarcinoma cell lines (Calu-3, A549 and A549-ACE2). Volcano plots of differential retroelement expression induced by related human coronaviruses in Calu-3 cells (A-D), various respiratory virus infections in A549 cells (E-I) and SARS-CoV-2 infection of A549-ACE2 cells at different MOI and pretreated with ruxolitinib kinase inhibitor (J-M).
Figure 2.

Network analysis of gene and retroelement expression changes induced by various viral infections. (A) Selected gene clusters produced from the comparison between the transcriptomes of Calu-3 cells in four infection conditions (MOCK; SARS_Cov2; SARS; MERS). For each cluster, the scaled mean expression value (z-score) for each gene in each group is plotted as a point. Box plots show the median expression value (dark line) and the first and third quartiles (edges) of the cluster’s gene expression values in each infection group. (B) On the left is shown a heatmap of Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway annotations enrichment for each cluster and (C) on the right the overrepresentation analysis of gene ontology (GO) for selected clusters. (D) Selected clusters produced from A549 cells and heatmaps of (E) KEGG pathways and (F) GO terms enriched for each cluster. (G) Selected clusters produced from A549-ACE2 cells and heatmaps of (H) KEGG pathways and (I) GO terms enriched for each cluster.
Figure 3.

ChIP-seq analysis of SARS-CoV-2 induced differentially expressed TE s. (A) Heatmap summarizing the transcription factor (TF) occupancy over individual differentially expressed (DE) TE proviral loci in fourteen human cell lines. This plot includes the distinct DE-TE loci that are occupied by at least one ChIP-seq peak. Each row represents an individual DE-TE locus. Each column shows TF ChIP-seq peak occupancy of a different cell line. Grey color denotes the absence, whereas dark red color denotes the presence of peaks in each locus. TFs are annotated on the top of the heatmap and showed in a legend on the right side. (B) Barplot shows the number of DE-TE loci occupied by each ChIP-seq in each cell line analyzed in Figure 3A. The bars are shown in ascending order, according to the number of enriched DE-TE loci. The ChIP-seq shown in the box is those that at least occupy half of the total DE-TE loci. (C) Heatmap shows the subsection of analysis in Figure 3A while only including the attested and predicted pioneer transcription factors. The color scheme of the plot and legends is the same as in Figure 3A. (D) This heatmap shows the subsection of analysis in Figure 3A only including the transcriptional and enhancer associated factors in A549 cells. The color scheme of the plot is the same as in Figure 3A. (E,F) Distribution of ChIP-seq signal of H3K27Ac around the DE-TE loci (7 kb upstream till 3 kb downstream sequences from the putative TSS) loci in HBTE (Human Bronchial Tracheal Epithelium) cells after the H1N1 infection at the time points viz. 6, 12, 18, 24 hours. (E) Genome browser tracks show the H3K27Ac signals at the specific set of loci shown in in A549 cells after SARS-CoV-2 infection at the time points viz. 0, 8, 24 hours.
Figure 4.

Differential retrotranscriptome profiling of RNA-seq data from clinical samples from COVID-19 donors. Volcano plots of differentially expressed TEs from BALF samples from COVID-19 donors compared to healthy donors separated by (A) HERVs and (B) LINE-1 elements. Volcano plots of differentially expressed TEs from PBMC of COVID-19 patients compared to healthy donors separated by (C) HERVs and (D) LINE-1 elements. (E) Upset plot of differentially expressed retroelement intersections across SARS-CoV-2 (COV2) infections of Calu-3 cells, A549 cells, A549-ACE2 cells, and BALF and PBMC from clinical samples of patients with severe COVID-19.
## Table Legends

<table>
<thead>
<tr>
<th>Cells</th>
<th>Infection</th>
<th>Concentration</th>
<th>Time after infection</th>
<th>Average % Viral Reads</th>
<th>Replicates</th>
<th>Treatment</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calu-3</td>
<td>SARS-CoV-2</td>
<td>2 MOI</td>
<td>24 hrs</td>
<td>15.05%</td>
<td>3</td>
<td>None</td>
<td>Blanco-Melo et al. (2020)</td>
</tr>
<tr>
<td>Calu-3</td>
<td>SARS-CoV</td>
<td>10^6 PFU/ml</td>
<td>24 hrs</td>
<td>57.71%</td>
<td>2</td>
<td>None</td>
<td>Yeung et al. (2016)</td>
</tr>
<tr>
<td>Calu-3</td>
<td>MERS-CoV</td>
<td>10^6 PFU/ml</td>
<td>24 hrs</td>
<td>48.15%</td>
<td>2</td>
<td>None</td>
<td>Yeung et al. (2016)</td>
</tr>
<tr>
<td>A549</td>
<td>SARS-CoV-2</td>
<td>2 MOI</td>
<td>24 hrs</td>
<td>0.10%</td>
<td>3</td>
<td>None</td>
<td>Blanco-Melo et al. (2020)</td>
</tr>
<tr>
<td>A549</td>
<td>IAV</td>
<td>5 MOI</td>
<td>9 hrs</td>
<td>10.80%</td>
<td>2</td>
<td>None</td>
<td>Blanco-Melo et al. (2020)</td>
</tr>
<tr>
<td>A549</td>
<td>RSV</td>
<td>2 MOI</td>
<td>24 hrs</td>
<td>3.97%</td>
<td>3</td>
<td>None</td>
<td>Blanco-Melo et al. (2020)</td>
</tr>
<tr>
<td>A549</td>
<td>HPIV3</td>
<td>3 MOI</td>
<td>24 hrs</td>
<td>6.71%</td>
<td>3</td>
<td>None</td>
<td>Blanco-Melo et al. (2020)</td>
</tr>
<tr>
<td>A549-ACE2</td>
<td>SARS-CoV-2</td>
<td>0.2 MOI</td>
<td>24 hrs</td>
<td>56.36%</td>
<td>3</td>
<td>None</td>
<td>Blanco-Melo et al. (2020)</td>
</tr>
<tr>
<td>A549-ACE2</td>
<td>SARS-CoV-2</td>
<td>2 MOI</td>
<td>24 hrs</td>
<td>62.25%</td>
<td>3</td>
<td>None</td>
<td>Blanco-Melo et al. (2020)</td>
</tr>
<tr>
<td>A549-ACE2</td>
<td>SARS-CoV-2</td>
<td>2 MOI</td>
<td>24 hrs</td>
<td>64.20%</td>
<td>3</td>
<td>Ruxolitinib</td>
<td>Blanco-Melo et al. (2020)</td>
</tr>
<tr>
<td>PBMC</td>
<td>SARS-CoV-2</td>
<td>N/A</td>
<td>N/A</td>
<td>1.31%</td>
<td>2</td>
<td>N/A</td>
<td>Xiong et al. (2020)</td>
</tr>
<tr>
<td>BALF</td>
<td>SARS-CoV-2</td>
<td>N/A</td>
<td>N/A</td>
<td>0.00%</td>
<td>3</td>
<td>N/A</td>
<td>Xiong et al. (2020)</td>
</tr>
</tbody>
</table>

Table 1.

Metadata of all RNA-seq samples accessed for analysis.