bioRxiv preprint doi: https://doi.org/10.1101/2021.10.17.464720; this version posted October 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Mapping Molecular Gene Signatures Among Respiratory Viruses Based on Large-Scale and Genome-wide Transcriptomics Analysis

Thomas Smith, Mohammed A. Rohaim, Muhammad Munir*

Division of Biomedical and Life Sciences, Lancaster University, Lancaster, Lancashire, LA1 4YG United Kingdom.

***Corresponding Author**

Dr Muhammad Munir

Lancaster University, Lancaster, United Kingdom.

Email: muhammad.munir@lancaster.ac.uk

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.17.464720; this version posted October 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is an emerging RNA virus causing COVID-19 disease across the globe. SARS-CoV-2 infected patients exhibit acute respiratory distress syndrome which can be compounded by endemic respiratory viruses and thus highlighting the need to understand the genetic bases of clinical outcome under multiple respiratory infections. In this study, 42 individual datasets and a multi-parametric based selected list of over 12,000 genes against five medically important respiratory viruses (SARS-CoV-2, SARS-CoV-1, influenza A, respiratory syncytial virus (RSV) and rhinovirus were collected and analysed in an attempt to understand differentially regulated gene patterns and to cast genetic markers of individual and multiple co-infections. While a certain cohort of virus-specific genes were regulated (negatively and positively), notably results revealed a greatest correlation among gene regulation by SARS-CoV-2 and RSV. Furthermore, out of analysed genes, the MAP2K5 and NFKBIL1 were specifically and highly upregulated in SARS-CoV-2 infection in vivo or in vitro. In contrast, several genes including GPBAR1 and SC5DL were specifically downregulated in SARS-CoV-2 datasets. Additionally, we catalogued a set of genes that were conserved or differentially regulated across all the respiratory viruses. These finding provide foundational and genome-wide data to gauge the markers of respiratory viral infections individually and under co-infection.

Key Words: SARS-CoV-2, SARS-CoV-1, Influenza, RSV, Rhinovirus, COVID-19, Transcriptomics.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.17.464720; this version posted October 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1. Background

Since its first appearance in Wuhan, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has rapidly spread across the world in a way unlike any other respiratory viruses. Coronavirus disease 2019 (COVID-19) caused by SARS-CoV-2 is considered the third highly pathogenic coronavirus following SARS-CoV-1 and Middle East respiratory syndrome coronavirus (MERS-CoV) that cause severe accurate respiratory symptoms in humans [1]. The most striking feature of the incidences and epidemiology of SARS-CoV-2 is its high ability for transmission among people [2]. The clinical outcome and incidence vary that most COVID-19 patients are clinically mild and moderate, and the elderly seem to have serious symptoms [3]. Additionally, severely affected patients had shown respiratory complications such as moderate to severe pneumonia, acute respiratory distress syndrome (ARDS), sepsis, acute lung injury (ALI), and multiple organ dysfunction (MOD) [4].

ARDS in COVID-19 patients is thought to be the main cause of death because of the cytokine storm caused by an over-activation of the human innate immune response [5]. However, there are multiple immune regulators and host genetic and epigenetic factors that are capable of significant contributions to the disease manifestation [5]. Host-pathogen interactions can act as a double-edged sword in different coronavirus infections as they might be useful not just for hosts, but also for viruses [6]. Similar tug-of-war host-viruses can also be present in COVID-19, which could lead to overcomplicated outcomes of the disease [7].

Although recent studies have shown the transcriptomic analysis of host responses to SARS-CoV-2 infection at different time points within multiple cell lines [8, 9], the transcriptional dynamics of host response to multiple virus infection remained largely unknown. In general, the host innate immune responses play an essential role in suppressing the replication of the virus once the virus enters the host, such as antiviral-mediated interferons and cytokines, which could lead to the virus pathogenesis. Increased cytokine levels are also observed in patients hospitalised with COVID-19 in the same way as both SARS-CoV and MERS-CoV, which induce high levels of cytokine [10, 11]. Accordingly, understanding the magnitude and dynamics of human transcriptome in response to medically important respiratory

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.17.464720; this version posted October 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

viruses will help in understanding their pathogenesis, molecular genetic markers and in repurposing existing antivirals to combat respiratory viral infections.

The current study aims to compare a large cohort of transcriptomic dataset map the gene regulation (up or down regulated) by SARS-CoV-2 infection and the compounding impact of other respiratory viruses such as influenza, SARS-CoV-1, respiratory syncytial virus (RSV) and rhinovirus. This parallel comparison showcases common and unique genetic signatures of respiratory viruses under individual and co-infection scenarios.

2. Materials and Methods

2.1 Data Collection, Inclusion and Exclusion Criteria

Gene Expression Omnibus (GEO) and PubMed datasets were used to search for literature that contained data relating to upregulated and downregulated genes in response to infection with respiratory viruses (SARS-CoV-2, influenza, SARS-CoV-1, RSV and rhinovirus). The collection began with searching for datasets for the more recent COVID-19 pandemic. On GEO, the terms *"("Severe acute respiratory syndrome coronavirus 2"[Organism] OR SARS-CoV-2[All Fields]) AND "Homo sapiens""* were used whereas when searching on PubMed, the terms "*(SARS-CoV-2) AND (Transcriptome)"* were used. Once datasets were identified, inclusion and exclusion criteria were carried out as outlined in **Table 1** to ensure parallel comparison of gene signatures.

2.2. Included Datasets and Data Synchronisation

The collected datasets from various sources were compiled into one set of data using Microsoft Excel program. The studied viruses and their respective analysed datasets are provided in a spreadsheet (**Table 2**). An overview of each dataset is provided in the **Supplementary dataset 1**. Each dataset carried genes found in a specific study mentioned in the category, and the corresponding level of gene expression is displayed next units originally used by the datasets. To ensure that all the included datasets for each virus could be compared, these were converted to the same units. The raw data was often listed in three units; Fold Change, Log Fold Change and Log 2-Fold Change, and all the data was converted into the Log 2-Fold Change format. Log 2-Fold Change was used as it allows easier visualisation of the

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.17.464720; this version posted October 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

data, as the range of the values of the data becomes narrower, allowing for easier comparison of the up/ down regulated genes **(Supplementary dataset 2).**

2.3. Ranking System

Owing to large diversity among datasets in areas such as cell types and media in which the experiments were carried out, it could introduce biasness to compare genes specifically by their Log 2-Fold Change values, which is calculated to the baseline gene expression. To introduce a novel method of comparing each gene up or down regulated in a dataset compared to datasets from another viruse or different cell types, a ranking system next to each Log 2-Fold Change column was proposed. This system ranked the genes based on which percentage group they were in, depending on whether they were up or down- regulated. Then, a mean score was taken across datasets within the same studied viruses and these means were used to compare between the viruses. For avoidance of confusion, this system synchronizes the dataset such that at the top 10% of upregulated genes for one virus while only at the top 80% of genes for another virus.

Using the GraphPad Prism 9.0.0 software, a scatter bar graph was generated using the overall ranking score for each gene of each virus. Two versions were created; first had the uncut data taken directly from the ranking system, containing roughly 24,000 genes, and secondly a cut down version of the data where non-significant genes were removed. Additionally, the non-coding gene loci and non-annotated genes were removed, as these often yielded zero values for up or down regulated genes reducing 6200 genes. Furthermore, other genes were removed which contained more than three or more zero values for up or downregulation across the five viruses removing a further 200 genes. Finally, using influenza virus as a model virus, all genes were removed that lied within the ranking scores of +20 (bottom 20% upregulated) and -20 (bottom 20% downregulated genes), unless a gene had a ranking score of above +50 or below -50 in any other of the viruses. This removed a further 5005 genes leaving a total of just over 12,000 genes in the cut down version, which removed the large proportion of genes containing zero values for clearer view for the spread of gene ranking scores.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.17.464720; this version posted October 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2.4. Log2 Fold Change

The collected dataset was converted into Log2 Fold Change for the gene expression. The datasets that were unconvertable into Log2 were removed for Log2 Fold Change analysis. In addition, only datasets that compared infected and non-infected patients were used while high vs low viral load datasets were removed. Finally, Log2 Fold Change values for each gene and for each dataset were inputted into the software and the graphing tool was used to generate scatter bar charts for each virus-specific dataset. These include the top five upregulated and/or downregulated genes for each dataset taken from the original data.

2.5. iDEP.91 Software

Once all the data had been converted into the ranking score format, it was exported into a separate Excel File to be compiled into one concise table **(Supplementary dataset 3)**, then saved as a text document and uploaded to the iDEP web application for expression and pathway analysis as described earlier [12].

3. Results

3.1. Overview of the differences in the Log-2 Fold Change values and Ranking Scores Across Multiple Respiratory Viruses

The scatter bar graphs for each of the individual datasets within each of the five viruses were drawn to provide an overview of the differences in the Log-2 Fold Change values obtained from each dataset (**Fig. 1)**. The scatter bar graph for the datasets collected for the SARS-CoV-2 uses the original Fold change values given by each study where each bar represents a separate dataset that showed the up and down regulated genes in response to viral infection (**Fig. 1A)**. A vast majority of top five upregulated genes were summarized (**Table 3**) while the top five down regulated genes involved in the innate immune response to SARS-CoV-2, SARS-CoV-1, influenza, RSV and rhinovirus infection were concluded (**Table 4**). Interestingly, each dataset shown was distinctive showing a varying pattern where host genes are mildly up or down regulated and only a few that are highly differentially up or down regulated. This highlights selective genes of the innate immune response are affected in response to a specific virus infection. Collectively, dataset GSE155286 has the widest spread of data while dataset GSE147507 has the lowest **(Table 2)**. In

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.17.464720; this version posted October 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

addition, all the datasets arried both downregulated and upregulated genes except GSE153790, which has only upregulated genes. Amongst the top five up-regulated genes, five genes including IFI27 and C-X-C motif chemokine ligand (CXCL) group of cytokine-producing genes, specifically CXCL10 showed a virus-specific trend.

Using the same approach, we used the data collected for innate immune genes in response to influenza virus infection that contains nine datasets. The data was presented for better visualising to gauge the innate immune genes play critical roles in the virus infection. Consistently, amongst all datasets, the up regulated genes for the influenza virus were interferon alpha-inducible protein 27 (IFI27) and interferon induced protein 44 producing gene IFI44/IFI44L, which involves in type-1 interferon signalling process leading to apoptosis and the formation of tubular structures, respectively.

The scatter bar graphs for SARS-CoV-1, RSV and rhinovirus indicate a unique set of genes up or down regulated during infection (**Fig. 1C, 1D and 1E),** respectively. While limited datasets were available against some viruses, minimum eight datasets provided approximately 12,000 different genes. Datasets that have gaps around the zero value for Log-2 fold change are the datasets that only include genes that were significantly up or down regulated. All datasets shown in **Fig. 1C, 1D and 1E** show a clear abundance of genes that are mildly differentially regulated with significantly less genes at the high fold change values, highlighted by the shape of the GSE53543 dataset. Interestingly, there was marked variation between the highest and lowest values obtained for log-2-fold change for different datasets within SARS-CoV-1. In addition, most of innate immune genes fall within +10 or -10 log-2 fold change for these viruses. However, SARS-CoV-1 appears to have a unique set of top five up regulated genes compared to the other viruses whereas both RSV and rhinovirus datasets showed IFI44 gene and the CXCL family. OASL remained a consistently upregulated gene in RSV datasets.

The log-2 fold change values of each gene for each dataset was changed into a ranking score due to the high variation of experimental method used to collect data for each dataset, which meant that log-2 fold change values were rarely consistent between datasets for differential gene regulation of patients/cells infected with the same respiratory virus. Thus, the ranking score removed this issue by assigning

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.17.464720; this version posted October 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

each gene a value based on its position among other differentially regulated genes within the same dataset (i.e., a gene placing as the 5th highest upregulated gene in a list of 100 genes would receive a score of 95). These synchronized values were averaged across all datasets within each virus that enabled the data collected from different experimental approaches to be compared more effectively between datasets within the same virus and a combination of datasets to be compared between different viruses **(Table 5 and 6)**.

3.2. iDEP.91 statistical analysis

The application of ranking scores facilitated the generation of a dataset consisting of 12,000 genes across all viruses by removing many non-significant genes (**Fig. 2A**). This newly and reduced set of genes and the data provided a higher resolution of genes distribution across multiple respiratory viruses (**Fig. 2B)**. Thereafter, all analysis was conducted using dataset generated through ranking system

The scatter plots generated on 12,000 genes highlight the distribution patterns of genes between SARS-CoV-2 and other respiratory viruses (**Fig. 2C to 2F)**. The relationship between SARS-CoV-2 and influenza virus gene regulation revealed a uniform scatter data (**Fig. 2C)**, while the relationship between SARS-CoV-2 and SARS-CoV-1 gene regulation contains more spread of data points except towards the centre of the graph due to the removal of less important data towards zero values **(Fig. 2D**). A slightly different patterns was observed where a linear relationship between SARS-CoV-2 and RSV (**Fig. 2E**) was noticed. An overall less uniform spread of data points with a skew to the right towards the top of the graph, and additional upregulated genes were observed in SARS-CoV-2 and rhinovirus comparison (**Fig. 2F)**.

3.3. Heatmap Analyses and Gene Differences between Respiratory Viruses

The heatmap were generated to provide an insight into pathways that are differently regulated by each of the five studied respiratory viruses (**Fig. 3)**. SARS-CoV-2 appeared unique in eliciting a separate viral response compared to the other respiratory viruses. Notably, there was a region at the bottom of the heatmap between genes DDX21 and GBP3 where other viruses have no effect or a slight upregulation of the genes, however, SARS-CoV-2 causes a downregulation (**Fig. 3)**.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.17.464720; this version posted October 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Perhaps the most unique out of all the respiratory viruses is SARS-CoV-1 which showed large areas of each heatmap where it is causing a downregulation of genes where all other respiratory viruses were eliciting upregulation.

The heatmap results highlighted the differences between each of the respiratory viruses, even though they are in the same group based on their target within the host; the genes that being affected are substantially different. Each virus shown in the heatmap carried different and distinct green and red areas, with very few coloured areas shared between more than two viruses. The most substantial difference was noticed between SARS-CoV-2 and SARS-CoV-1, whereas almost no colours in common. However, SARS-CoV-1 appeared to be the only virus that has both up and down regulated genes in two specific groups.

3.4. Standard deviation calculation and T-SNE plot Analyses

The SD graph highlights the extremely high standard deviation across all the regulated genes in response to different viruses (**Fig. 4A)**. A standard deviation above 1 was considered high unless the standard deviation in this case was between 25 and 75 indicating that there are high differences in the differentially regulated genes in response to each virus. On the other hand, a correlation matrix that shows the correlation between each of the viruses revealed that the most similar virus to SARS-CoV-2 was RSV with a Pearson's correlation coefficient of 0.48 **(Fig. 4B)** while the least similar one was SARS-CoV-1 with a Pearson's correlation coefficient of 0.15 **(Fig. 4B)**. A correlation value of 1 implies that there is a perfectly linear distribution of data between the two variables and a value of 0.48 generated for RSV compared to SARS-CoV-2 is relatively high that highlight how close the two viruses are in comparison to other viruses.

Differentially regulated genes were classified into 20 clusters based on their K means (**Fig. 4C)** where we used them to break down for better understanding whereabouts the differences between these emerged viruses. Each cluster contains genes involved in specific pathways that allows for the comparison of gene regulation in a variety of pathways depending on the virus **(Supplementary dataset 4)**. After K-means clustering, cluster O appeared to contain the most pathways involved in the innate immune response, such as the JAK-STAT signalling pathway, TNF signalling pathway and IL-17 signalling pathway indicating that cluster O could

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.17.464720; this version posted October 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

be used as a sign of a virus's regulation for the overall innate immune response signalling. Both influenza and SARS-CoV-2 showed both up and down regulated genes within the cluster with specific areas either being highly up- or down-regulated, suggesting that these viruses target specific areas within this cluster. While SARS-CoV-1 and RSV upregulated and down regulated this region, respectively.

The T-SNE plot analyses for all the data was coloured based on their belonging cluster. The T-SNE allowed multi-dimensional data to visualise in a low dimension space such as the 2D graph **(Fig. 4D)**. The distance between each of the points reflected the similarity of each data point. Whilst T-SNE should not always be used for gene expression data analysis, due to its high intrinsic dimensionality. Therefore, it has been used to highlight that even though there are a high number of clusters present, they are still very much distinguishable, despite there being some clusters that exhibit more separation of data points compared to others. In addition, there was a slight problem with crowding towards the centre of the dataset; however, this was observed in most SNE forms.

3.5. Comparison between differentially regulated genes among multiple respiratory viruses

Generally, the number of upregulated genes is relatively even with the number of downregulated genes, however, there are more downregulated genes than upregulated genes for each of the five tested viruses. The standouts are substantially downregulated than upregulated genes in case of rhinovirus infection in (**Fig. 5A)**. Moreover, rhinovirus showed less differentially regulated genes in total compared to the other respiratory viruses.

The Venn diagrams showed a comparison between each of the viruses by how many differentially regulated genes they have in common, regardless of whether they are up or down regulated. This highlighted genes that are differently regulated within only one virus compared to others within the same diagram (**Fig. 5B**). Vast majority of genes are found to be differentially expressed across all viruses; however, there were some exceptions mainly found within RSV that has the highest number of genes unique to itself while rhinovirus rarely had any uniquely expressed genes.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.17.464720; this version posted October 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3.6. Impact of SARS-CoV-2 on cellular DNA replication

A visual representation for the impact of SARS-CoV-2 on the DNA replication within infected cells was outlined (**Fig. 6)**. The upregulated genes (2/32) were shown in red while the downregulated (27/32) were shown in green **(Fig. 6)**. Genes responsible to produce DNA ligase and helicase were notably down regulated, which are important in the DNA replication and being used by the virus as a means of slowing down the cell cycle to enhance viral replication.

3.7. Regulation of JAK-STAT immune signalling pathway in response to SARS CoV-2 infection

There are more upregulated genes in JAK-STAT immune signalling (**Fig. 7A)** and the cytokine-cytokine receptor interaction pathways (**Fig. 7B)** than downregulation, highlighted by the prominence of the red colouring over the green colouring. While there were several upregulated genes as GFAP and Ras, which are involved in cell differentiation and MAPK signalling pathway, respectively.

The genes that are up or down regulated in realtion to the immune signalling pathways and are affected in response to SARS-CoV-2 infection were analysed using KEGG pathway database (**Fig. 7A, 7B and Supplematary Figure 1)**. These results revealed that SARS-CoV-2 does not affect every pathway in a simple manner by either upregulating or downregulating all genes involved in that pathway, but instead having multiple effects.

Using the ranking scores, C8orf4 was the second most highly upregulated gene in cells/patients infected with SARS-CoV-2. The C8orf4 (also known as TCIM) is responsible for producing the c8orf4 protein (also known as TC1) which is involved in the enhancement of NF-kappaB activity, leading to up-regulating several cytokines involved in the process of inflammation [13]**.** This is the main factor attributed to the cytokine storm exhibited in patients following SARS-CoV-2 infection. In addition, our analyses show that each virus has a different effect on the regulation of c8orf4 and its regulation could therefore be used as a biomarker to differentiate between aetiology of infection, with extremely high levels of TC1 protein pointing towards a SARS-CoV-2 infection **(Fig. 7C)**. Of course, many other genes could be used as

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.17.464720; this version posted October 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

markers for SARS-CoV-2 infection but also genes that are conserved between all viruses.

After individual identification of the upregulated or downregulated genes and their respective pathways, we aim to visualise where those genes are located within the human chromosome **(Fig. 7D)**. Human genome map analyses show each chromosome with its own line with genes where the upregulated genes appear above the line in red colour while genes that are downregulated appear below the line in blue colour **(Fig. 7D)**. This genomic map shows the regulation in response to SARS-CoV-2 infection and revealed that every chromosome in the human genome has been affected whereas the mostly affected chromosome was chromosome 19. However, the least affected chromosomes were X and Y sex chromosomes. In addition, chromosome 17 also shows a notable pattern. There are many areas across many chromosomes that showed notable gaps where SARS-CoV-2 appears to have no effect on gene regulation **(Fig. 7D)**.

There is a large amount of consistency between all the genome maps within the most affected chromosome, in all cases, being chromosome 19. In case of rhinovirus, there is a lack of altered genes regulation on the X and Y chromosomes. Furthermore, a much blander overall picture on fewer data points (**Supplementary figure 2D**) because there were less genes recorded to have been up or downregulated in the rhinovirus dataset.

4. Discussion

Despite majority of the human respiratory viruses show similar pathology by infecting the same respiratory system, they all showed clear and substantial differences, which have highlighted unique markers related to differential gene regulation. The scatter plots showed the correlation between the effects of each virus on human gene expression, and a specific removal of genes was evident in this analysis which are less dramatically differentially regulated and therefore of less importance to this study. These results indicated that SARS-CoV-2 is like RSV compared to other respiratory viruses because of the high correlation between the data points within the scatter graphs showing a rising diagonal line suggesting a positive correlation

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.17.464720; this version posted October 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

between the upregulated and downregulated genes. These results are supported by the correlation matrix, where the Pearson's correlation coefficient between SARS-CoV-2 and RSV was 0.48, much higher than the 0.22, 0.2 and 0.15 for influenza, rhinovirus, and SARS-CoV-1, respectively. The SARS-CoV-2 and RSV showed high similarity in differentially regulated genes. This aligns with the fact that affected patients exhibit similar symptoms when infected with any of SARS-CoV-2 or RSV, mainly upper respiratory tract symptoms and often lower respiratory tract symptoms such as a dry cough [14]. Interestingly, both viruses appear to cause damage to the respiratory tract that result in persistent problems long after infection such as persistent airway obstruction as well as hyper-responsiveness can be seen in patients 30 years after infection with RSV [15]**.** These symptoms are like the long-term lung dysfunction reported after SARS-CoV-2 infection [16]**.** However, the main difference between these two viruses is the age of the patients that are more susceptible for infection, with RSV commonly causing respiratory tract infection in young infants and children [14], whereas SARS-CoV-2 is known for more severe cases being present in the elderly albeit infection potential among all ages. Further research in this area could be useful to compare influenza, RSV, SARS-CoV-1 and rhinovirus against SARS-CoV-2 but specifically for each pathway/area such as the innate immune response or the cytokine activation pathway.

Insights into the human chromosomes in response to SARS-CoV-2 infection revealed that the mostly affected chromosome was chromosome 19 suggesting a high number of genes involved in the immune response to viral infection could be present within chromosome 19 and severe cases of infection could be attributed to the genetic mutations within this chromosome. Another interesting point is the presence of differential gene expression on the X chromosome for patients suffering from COVID-19. Altered genes on the X-chromosome could lead to a difference in the clinical outcome between men and women infected with SARS-CoV-2. Previous studies reported that the immune regulatory genes encoded by the X chromosome in women could cause lower viral load levels resulting in a reduction in the inflammatory response compared to men [17]**.**

The top and bottom five consistently up and down- regulated genes across all five viruses could potentially be used as markers for specific respiratory viral infection. JAK2 is one of the genes, which is consistently, and highly upregulated among all

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.17.464720; this version posted October 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the studied viruses and it encodes for the Janus Kinase 2 protein (JAK2). JAK2 plays a crucial role in the cytokine signalling where it associates with type II cytokine receptors, hormone-like cytokine receptors and being activated by IFN-gamma [18]**.** Additional four-upregulated genes were DDX60L, IFI44, FOXN2 and DDX60, which may be a target for drugs.

The upregulated genes in response to SARS-CoV-2 infection have been identified while those were downregulated in the other respiratory viruses. These genes could be used as markers for a SARS-CoV-2 infection and to distinguish SARS-CoV-2 from other respiratory viruses. The most important gene was the NFKBIL1 gene that encodes for the NF-kappa-B inhibitor-like protein 1 and it is involved in the NF-kappa-B signalling, which plays a major role in the inflammatory response by increasing the cytokine expression [19]**.** On the other hand, the downregulated genes in response to SARS-CoV-2 could possibly be used as a marker to distinguish SARS-CoV-2 infection in case of suspicion with a respiratory virus infection associated with respiratory symptoms. One of these genes is GPBAR1, which encodes for the G-protein acid receptor 1. Previously, it has been reported that GPBAR1 was able to regulate and increase the expression of IL-10 [20] suggesting that levels of IL-10 in patients suffering with COVID-19 would be lower, however, recent studies contradict that as IL-10 levels are found to be unexpectedly increased in severe cases [21]**.**

5. Conclusions and limitations

The aim of this study was to determine the influence of SARS-COV-2 on the immune regulation and gene induction in comparison to other respiratory viruses. It appeared that SARS-CoV-2 was unique in its impact on gene regulation and matches none of the other respiratory viruses except RSV. Genes such as MAP2K5 and NFKBIL1 have been found to be greatly upregulated in SARS-CoV-2 whilst being downregulated in the compared viruses. Whereas genes such as GPBAR1 and SC5DL were contrastingly found to be significantly downregulated in SARS-CoV-2 but upregulated in influenza, SARS-CoV-1, RSV and rhinovirus. Despite all the reported differences, the most conserved genetic signature was JAK2 gene as well as the constitutively downregulated ZNF219 gene. While the resolution of analysis

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.17.464720; this version posted October 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

provides foundational finding, further research is warranted to validate the impact of these molecular signature against individual or multiple infections.

We observed a limitation of the study that the gene regulation may be affected by the experimental characteristics such as time length post infection, the culturing conditions, phenotypes of the cells, and the nature of the virus stimulation (in vivo or in vitro studies). Finally, different cell types (A549, BALF or PBMC cells) were used for virus infection, which may respond differently to different viral infections. Nevertheless, the provided analysis provides a foundation for the impact of respiratory viruses on the gene regulation.

Ethics Statement

No experimental procedures were carried out in this project and all data was collected from a range of previous research papers; therefore, no steps of this study were required to seek ethical approval.

Declaration of competing interest.

The author has declared that no competing interests exist.

References

[1] K. Sun, J. Chen, C. Viboud, Early epidemiological analysis of the coronavirus disease 2019 outbreak based on crowdsourced data: a population-level observational study. *The Lancet. Digital health*, *2* (2020) e201–e208. https://doi.org/10.1016/S2589-7500(20)30026-1

[2] Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, et al., Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. N. Engl. J. Med. 382 (2020), 1199–207. https://doi.org/10.1056/NEJMoa2001316

[3] W.J. Guan, Z.Y. Ni, Y. Hu, W.H. Liang, C.Q. Ou, et al., Clinical Characteristics of Coronavirus Disease 2019 in China. N. Engl. J. Med. 382 (2020), 1708-1720. https://doi.org/10.1056/NEJMoa2002032

[4] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, et al., Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. 395 (2020), 497-506. https://doi.org/10.1016/S0140-6736(20)30183-5

[5] J.P. Salameh, M.M. Leeflang, L. Hooft, N. Islam, T.A. McGrath, C.B. van der Pol, et al., Cochrane COVID-19 Diagnostic Test Accuracy Group, McInnes MD. Thoracic

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.17.464720; this version posted October 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

imaging tests for the diagnosis of COVID-19. Cochrane Database Syst .Rev. 9 (2020), CD013639. https://doi.org/10.1002/14651858.CD013639.pub2

[6] T.S. Fung, D.X. Liu, Human Coronavirus: Host-Pathogen Interaction. Annu. Rev. Microbiol. 73 (2019), 529-557. https://doi.org/10.1146/annurev-micro-020518-115759.

[7] D.E. Gordon, G.M. Jang, M. Bouhaddou, J. Xu, K. Obernier, K.M. White et al., A SARS- CoV-2 protein interaction map reveals targets for drug repurposing. Nature. 583 (2020), 459–468. https://doi.org/10.1038/s41586-020-2286-9

[8] D. Blanco-Melo, B.E. Nilsson-Payant, W.C. Liu, S. Uhl, D. Hoagland et al., Imbalanced Host Response to SARS-CoV-2 Drives Development of COVID-19. Cell. 181(2020),1036-1045.e9. https://doi.org/10.1016/j.cell.2020.04.026.

[9] J. Xiong, O. Lipsitz, F. Nasri, L. Lui, H. Gill, L. Phan, D. Chen-Li et al., Impact of COVID-19 pandemic on mental health in the general population: A systematic review. J. Affect. Disord. 277 (2020), 55–64. https://doi.org/10.1016/j.jad.2020.08.001

[10] R. Channappanavar, S. Perlman, Pathogenic human coronavirus infections: causes and consequences of cytokine storm and immunopathology. Semin Immunopathol. 39 (2017), 529-539. https://doi.org/10.1007/s00281-017-0629-x.

[11] Z. Song, Y. Xu, L. Bao, L. Zhang, P. Yu, Y. Qu, H. Zhu, W. Zhao, Y. Han, C. Qin, From SARS to MERS, Thrusting Coronaviruses into the Spotlight. Viruses. 11 (2019), 59. https://doi.org/10.3390/v11010059.

[12] S.X. Ge, E.W. Son, R. Yao, iDEP: an integrated web application for differential expression and pathway analysis of RNA-Seq data. BMC bioinformatics, *19* (2018), 534. https://doi.org/10.1186/s12859-018-2486-6

[13] J. Kim, Y. Kim, H.T. Kim, D.W. Kim, Y. Ha, J. Kim, C.H. Kim, I. Lee, K. Song, TC1(C8orf4) is a novel endothelial inflammatory regulator enhancing NF-kappaB activity. J Immunol. 183 (2009), 3996-4002. https://doi.org/10.4049/jimmunol.0900956.

[14] D.K. Smith, S. Seales, C. Budzik, Respiratory Syncytial Virus Bronchiolitis in Children. American family physician. 95 (2017), 94–99.

[15] E. Priante, M.E. Cavicchiolo, E. Baraldi, RSV infection and respiratory sequelae. Minerva pediatrica. 70 (2018), 623–633. https://doi.org/10.23736/S0026-4946.18.05327-6

[16] Centers for Disease Control and Prevention. COVID-19 And Your Health. [online] Available at: https://www.cdc.gov/coronavirus/2019-ncov/long-term-effects.html [Accessed 24 January 2021].

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.17.464720; this version posted October 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[17] P. Conti, A. Younes, Coronavirus COV-19/SARS-CoV-2 affects women less than men: clinical response to viral infection. J. Biol. Regul. Homeost. 34 (2020), 339–343. https://doi.org/10.23812/Editorial-Conti-3

[18] K. Ghoreschi, A. Laurence, J.J. O'Shea, Janus kinases in immune cell signaling. Immunological reviews, 228 (2009), 273–287. https://doi.org/10.1111/j.1600-065X.2008.00754.x

[19] T. Liu, L. Zhang, D. Joo, S.C. Sun, NF-κB signaling in inflammation. Signal transduction and targeted therapy, 2 (2017), 17023. https://doi.org/10.1038/sigtrans.2017.23

[20] M. Biagioli, A. Carino, S. Cipriani, D. Francisci, S. Marchianò, P. Scarpelli, D. Sorcini, A. Zampella, S. Fiorucci, The Bile Acid Receptor GPBAR1 Regulates the M1/M2 Phenotype of Intestinal Macrophages and Activation of GPBAR1 Rescues Mice from Murine Colitis. J. Immunol. 199 (2017), 718-733. https://doi.org/10.4049/jimmunol.1700183

[21] L. Lu, H. Zhang, D. Dauphars, Y. He, 2021. A Potential Role of Interleukin 10 in COVID-19 Pathogenesis. Trends Immunol. 42 (2021), 3-5. https://doi.org/10.1016/j.it.2020.10.012

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.17.464720; this version posted October 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure captions

Fig. 1. Scatter bar graphs of the Log-2-Fold Change of each gene for each dataset for the A) SARS-CoV-2, B) Influenza, C) SARS-CoV-1, D) RSV, E) Rhinovirus. A horizontal line is also shown on each bar, which marks the average Log-2 fold change of the selected genes.

Fig. 2. Uncut (A) and Cut (B) ranking scores for each gene combining all datasets for each respiratory virus. Also, in this figure are scatter plots of ranking scores of all genes collected for each respiratory virus, using SARS-CoV-2 as the comparison. (C) shows a comparison of Influenza and SARS-CoV-2, (D) between SARS-CoV-1 and SARS-CoV-2, (E) between RSV and SARS-CoV-2 and (F) between Rhinovirus and SARS-CoV-2.

Fig. 3. Heatmap of DEGs for all the respiratory viruses studied in this analysis.

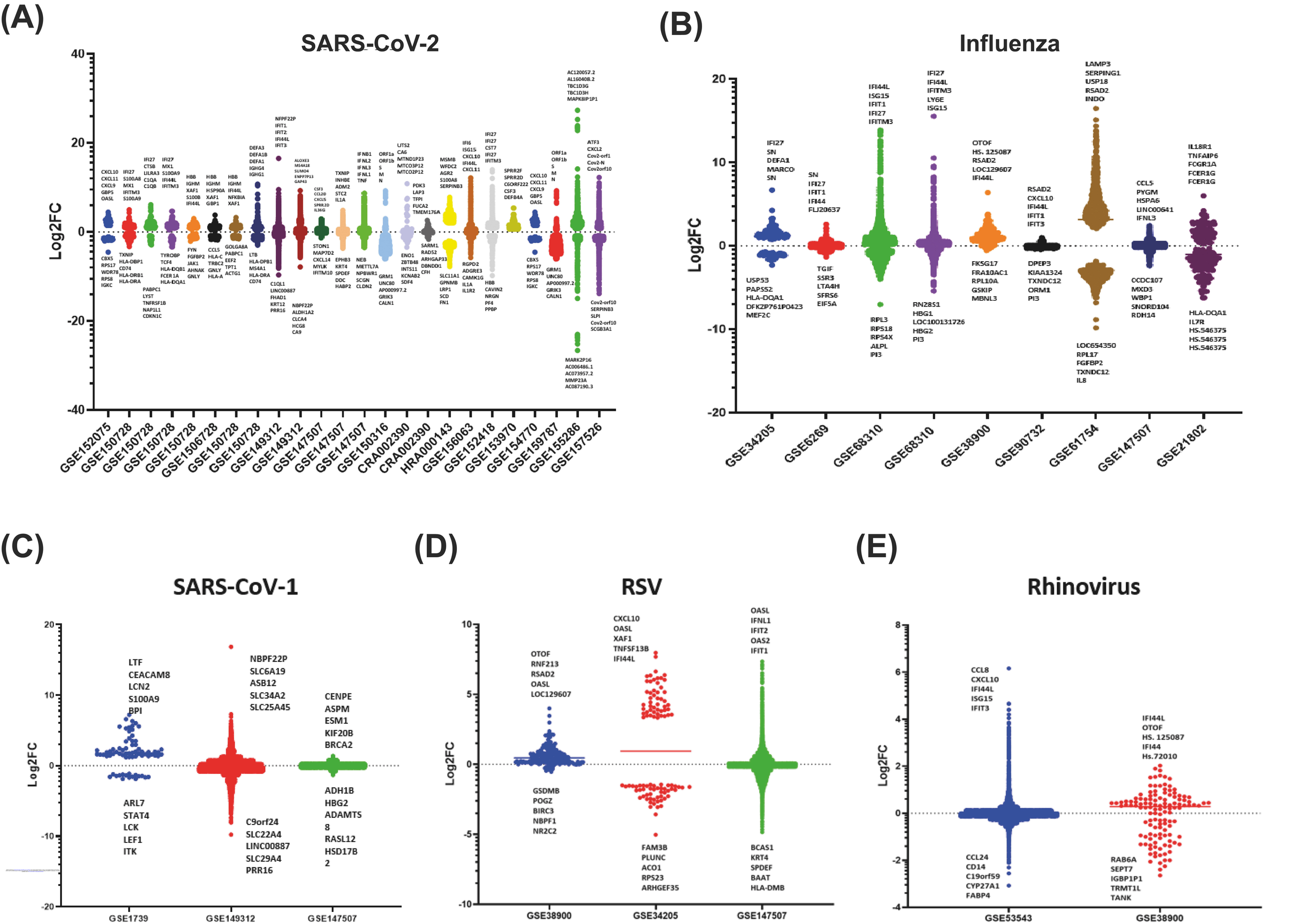
Figure 4. (A) Standard deviation of all genes across all viruses. (B) Correlation matrix using data taken from the top 75% of genes. (C) KEGG pathway analysis by cluster. (D) T-SNE plot of all 12,000 genes

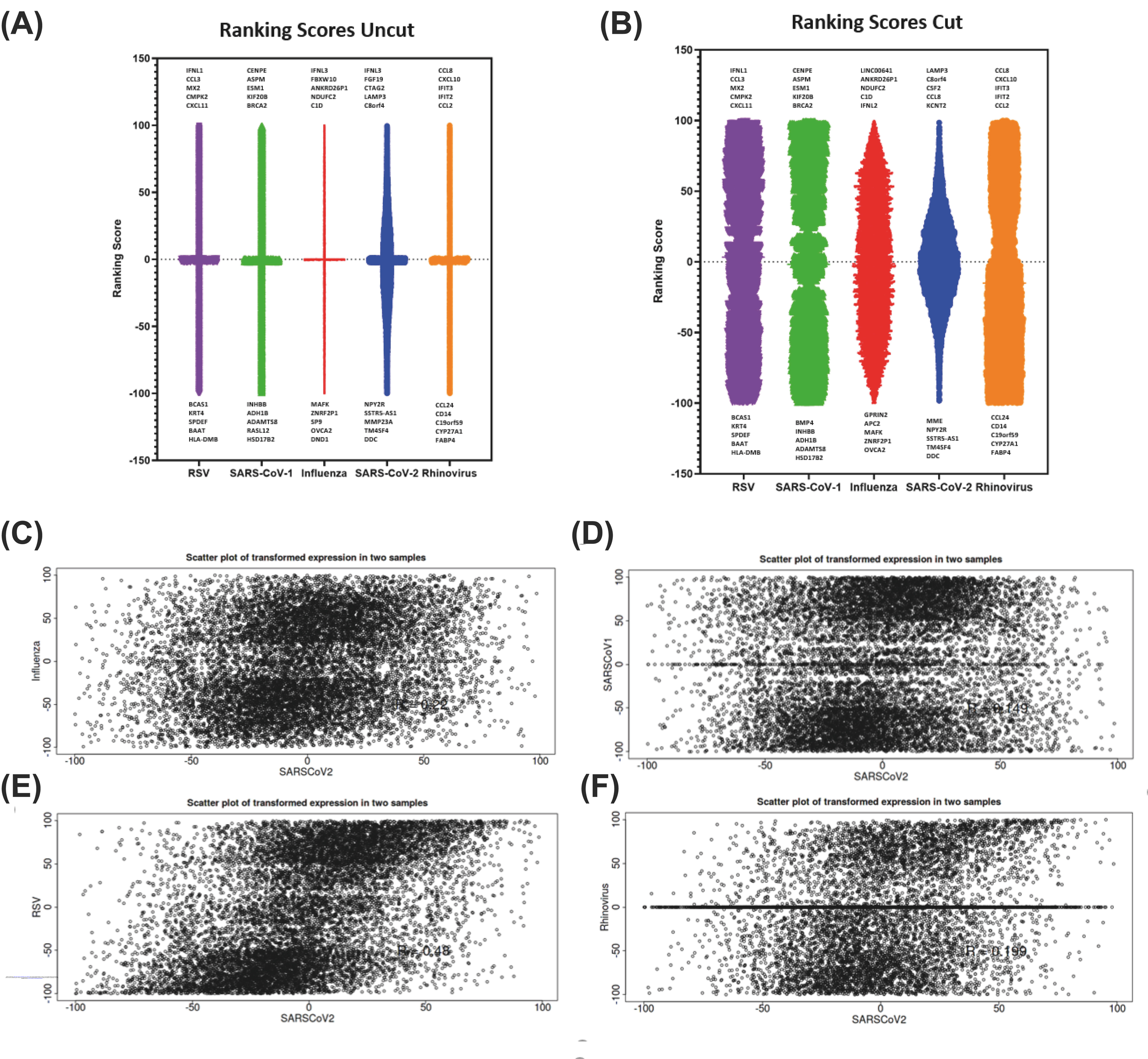
Fig. 5. (A) Total number of upregulated and down regulated genes for each virus. (B) Venn diagrams representing the differentially regulated genes that are in common between each of the respiratory viruses.

Fig. 6. Heatmaps specific to different pathways compiled by GAGE pathway analysis. (A) for Defence response to virus, (B) for cytokine response, (C) for regulation of cytokine production and (D) for positive regulation of innate immune response.

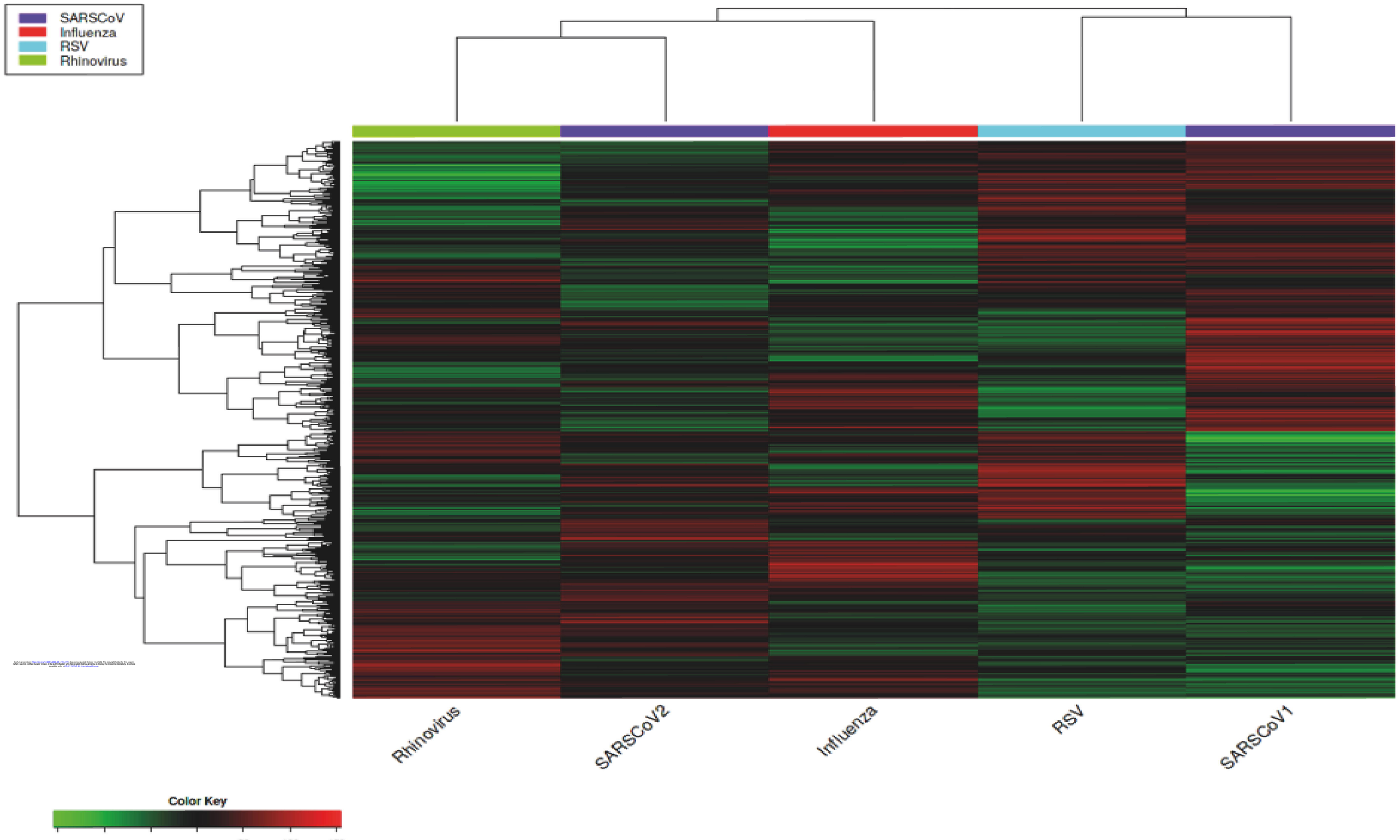
Fig. 7. (A) Regulation of genes associated with the JAK-STAT signalling pathway. (B) Regulation of genes associated with cytokine-cytokine receptor interaction. (C) Ranking scores of the C8orf4 gene for each respiratory virus. (D) Genome map showing SARS-CoV-2 upregulated genes in red and downregulated genes in blue.

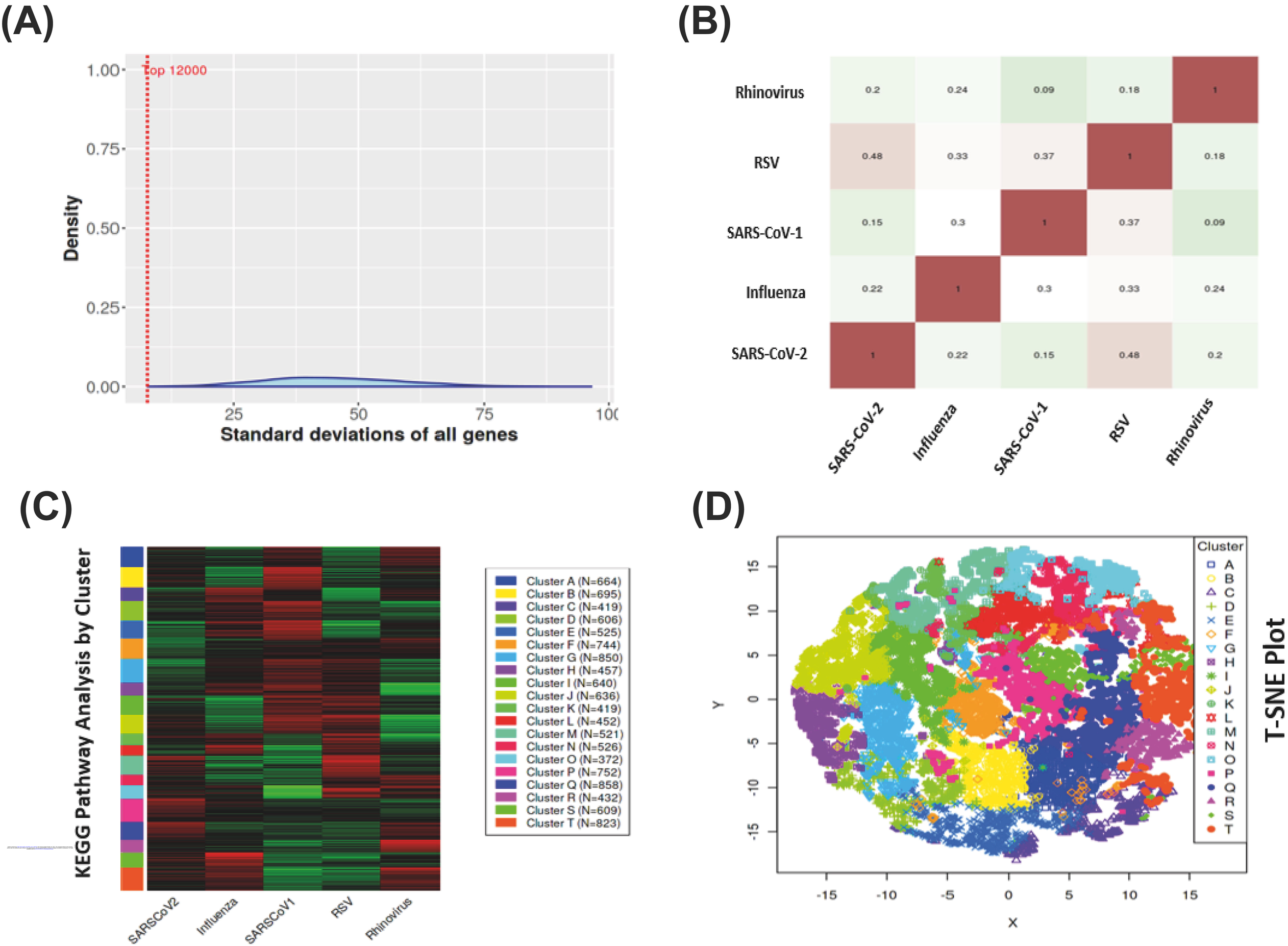
bioRxiv preprint doi: https://doi.org/10.1101/2021.10.17.464720; this version posted October 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

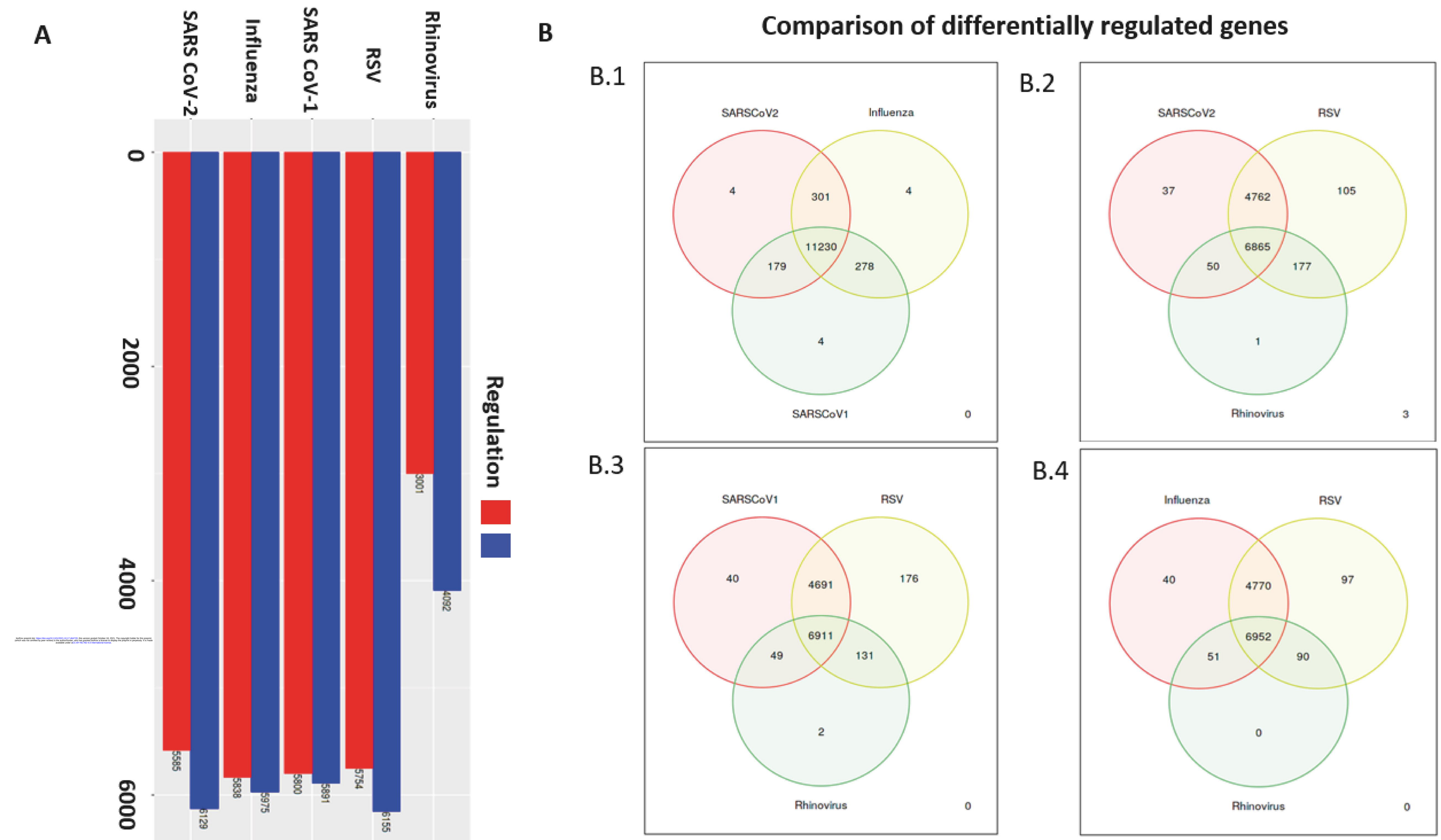

(A)
SARS-CoV-2
Log2FC
40
20
0
-20
-40
GSE152075
GSE150728
GSE150728
GSE150728
GSE150728
GSE1506728
GSE150728
GSE150728
GSE149312
GSE149312
GSE147507
GSE147507
GSE147507
GSE150316
CRA002390
CRA002390
HRA000143
GSE156063
GSE152418
GSE153970
GSE154770
GSE159787
GSE155286
GSE157526
(B)
Influenza
Log2FC
20
10
0
-10
-20
GSE34205
GSE6269
GSE68310
GSE68310
GSE38900
GSE90732
GSE61754
GSE147507
GSE21802
(C)
SARS-CoV-1
Log2FC
20
10
0
-10
-20
LTF
CEACAM8
LCN2
S100A9
BPI
ARL7
STAT4
LCK
LEF1
ITK
NBPF22P
SLC6A19
ASB12
SLC34A2
SLC25A45
C9orf24
SLC22A4
LINC00887
SLC29A4
PRR16
CENPE
ASPM
ESM1
KIF20B
BRCA2
ADH1B
HBG2
ADAMTS
8
RASL12
HSD17B
2
GSE1739
GSE149312
GSE147507
(D)
RSV
Log2FC
10
5
0
-5
-10
OTOF
RNF213
RSAD2
OASL
LOC129607
GSDMB
POGZ
BIRC3
NBPF1
NR2C2
CXCL10
OASL
XAF1
TNFSF13B
IFI44L
FAM3B
PLUNC
ACO1
RPS23
ARHGEF35
OASL
IFNL1
IFIT2
OAS2
IFIT1
BCAS1
KRT4
SPDEF
BAAT
HLA-DMB
GSE38900
GSE34205
GSE147507
(E)
Rhinovirus
Log2FC
8
6
4
2
0
-2
-4
CCL8
CXCL10
IFI44L
ISG15
IFIT3
CCL24
CD14
C19orf59
CYP27A1
FABP4
IFI44L
OTOF
HS. 125087
IFI44
Hs.72010
RAB6A
SEPT7
IGBP1P1
TRMT1L
TANK
GSE53543
GSE38900

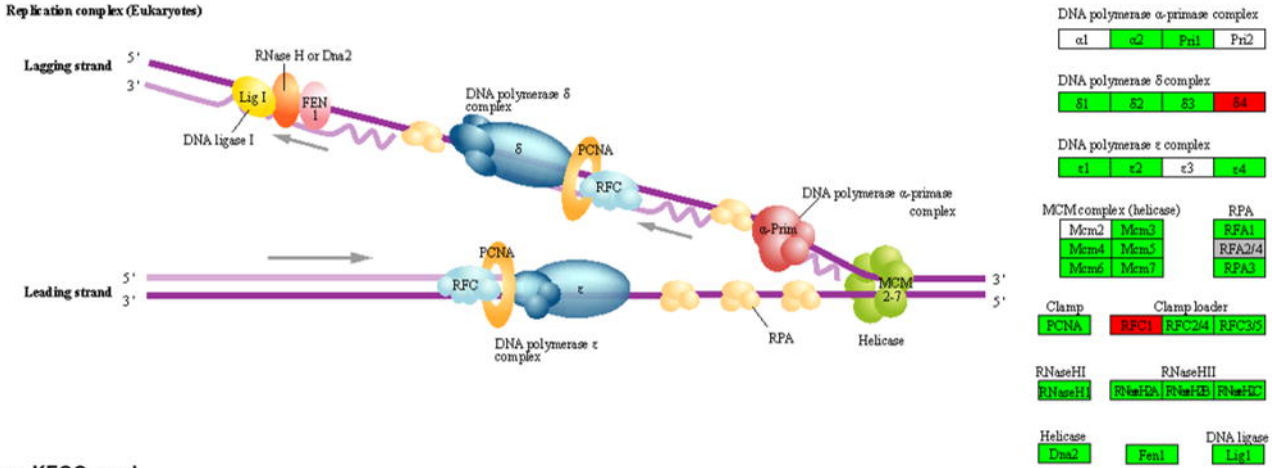

(A)
Ranking Scores Uncut
Ranking Score
150
100
50
0
-50
-100
-150
IFNL1 CCL3 MX2 CMPK2 CXCL11
CENPE ASPM ESM1 KIF20B BRCA2
IFNL3 FBXW10 ANKRD26P1 NDUFC2 C1D
IFNL3 FGF19 CTAG2 LAMP3 C8orf4
CCL8 CXCL10 IFIT3 IFIT2 CCL2
BCAS1 KRT4 SPDEF BAAT HLA-DMB
INHBB ADH1B ADAMTS8 RASL12 HSD17B2
MAFK ZNRF2P1 SP9 OVCA2 DND1
NPY2R SSTR5-AS1 MMP23A TM4SF4 DDC
CCL24 CD14 C19orf59 CYP27A1 FABP4
RSV
SARS-CoV-1
Influenza
SARS-CoV-2
Rhinovirus
(B)
Ranking Scores Cut
Ranking Score
150
100
50
0
-50
-100
-150
IFNL1 CCL3 MX2 CMPK2 CXCL11
CENPE ASPM ESM1 KIF20B BRCA2
LINC00641 ANKRD26P1 NDUFC2 C1D IFNL2
LAMP3 C8orf4 CSF2 CCL8 KCNT2
CCL8 CXCL10 IFIT3 IFIT2 CCL2
BCAS1 KRT4 SPDEF BAAT HLA-DMB
BMP4 INHBB ADH1B ADAMTS8 HSD17B2
GPRIN2 APC2 MAFK ZNRF2P1 OVCA2
MME NPY2R SSTR5-AS1 TM4SF4 DDC
CCL24 CD14 C19orf59 CYP27A1 FABP4
RSV
SARS-CoV-1
Influenza
SARS-CoV-2
Rhinovirus
(C)
Scatter plot of transformed expression in two samples
Influenza
SARSCoV2
R = 0.22
(D)
Scatter plot of transformed expression in two samples
SARSCoV1
SARSCoV2
R = 0.149
(E)
Scatter plot of transformed expression in two samples
RSV
SARSCoV2
R = 0.48
(F)
Scatter plot of transformed expression in two samples
Rhinovirus
SARSCoV2
R = 0.199

Heatmap


SARSCoV
Influenza
RSV
Rhinovirus
Rhinovirus
SARSCoV2
Influenza
RSV
SARSCoV1
Color Key
-150
-100
-50
0
50
100
150

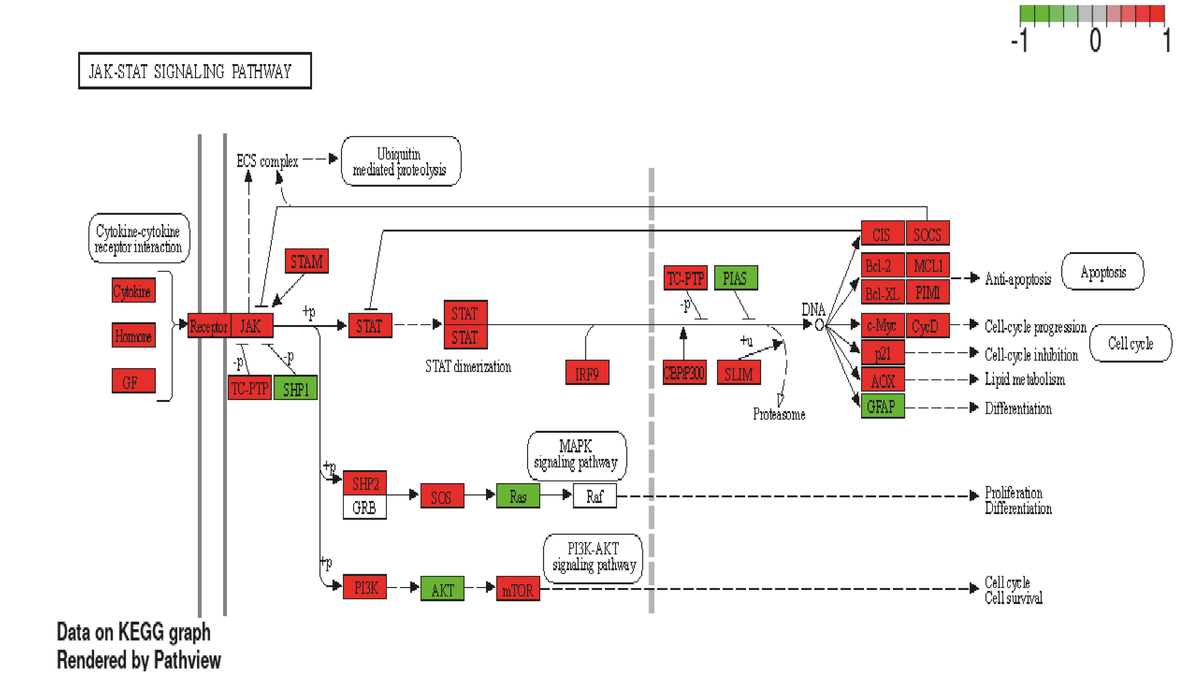

(A)
Top 12000
Density
1.00
0.75
0.50
0.25
0.00
25
50
75
100
Standard deviations of all genes
(B)
Rhinovirus
RSV
SARS-CoV-1
Influenza
SARS-CoV-2
0.2 0.24 0.09 0.18 1
0.48 0.33 0.37 1 0.18
0.15 0.3 1 0.37 0.09
0.22 1 0.3 0.33 0.24
1 0.22 0.15 0.48 0.2
SARS-CoV-2
Influenza
SARS-CoV-1
RSV
Rhinovirus
(C)
KEGG Pathway Analysis by Cluster
SARSCoV2
Influenza
SARSCoV1
RSV
Rhinovirus
Cluster A (N=664)
Cluster B (N=695)
Cluster C (N=419)
Cluster D (N=606)
Cluster E (N=525)
Cluster F (N=744)
Cluster G (N=850)
Cluster H (N=457)
Cluster I (N=640)
Cluster J (N=636)
Cluster K (N=419)
Cluster L (N=452)
Cluster M (N=521)
Cluster N (N=526)
Cluster O (N=372)
Cluster P (N=752)
Cluster Q (N=858)
Cluster R (N=432)
Cluster S (N=609)
Cluster T (N=823)
(D)
Cluster
A B C D E F G H I J K L M N O P Q R S T
Y
X
-15 -10 -5 0 5 10 15
T-SNE Plot


A
SARS CoV-2
Influenza
SARS CoV-1
RSV
Rhinovirus
0
2000
4000
6000
5585
6129
5838
5975
5800
5891
5754
6155
3001
4092
Regulation
B
Comparison of differentially regulated genes
B.1
SARSCoV2
Influenza
SARSCoV1
4
301
4
11230
179
278
4
0
B.2
SARSCoV2
RSV
Rhinovirus
37
4762
105
6865
50
177
1
3
B.3
SARSCoV1
RSV
Rhinovirus
40
4691
176
6911
49
131
2
0
B.4
Influenza
RSV
Rhinovirus
40
4770
97
6952
51
90
0
0



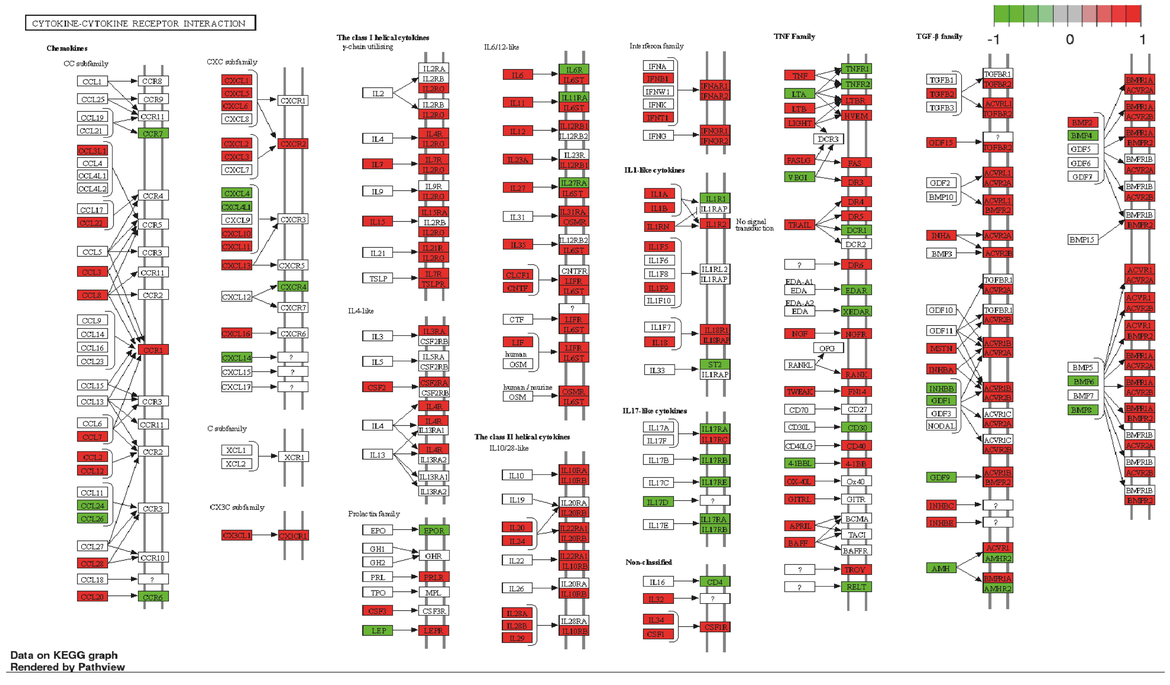
Data on KEGG graph
Rendered by Pathview

(A)



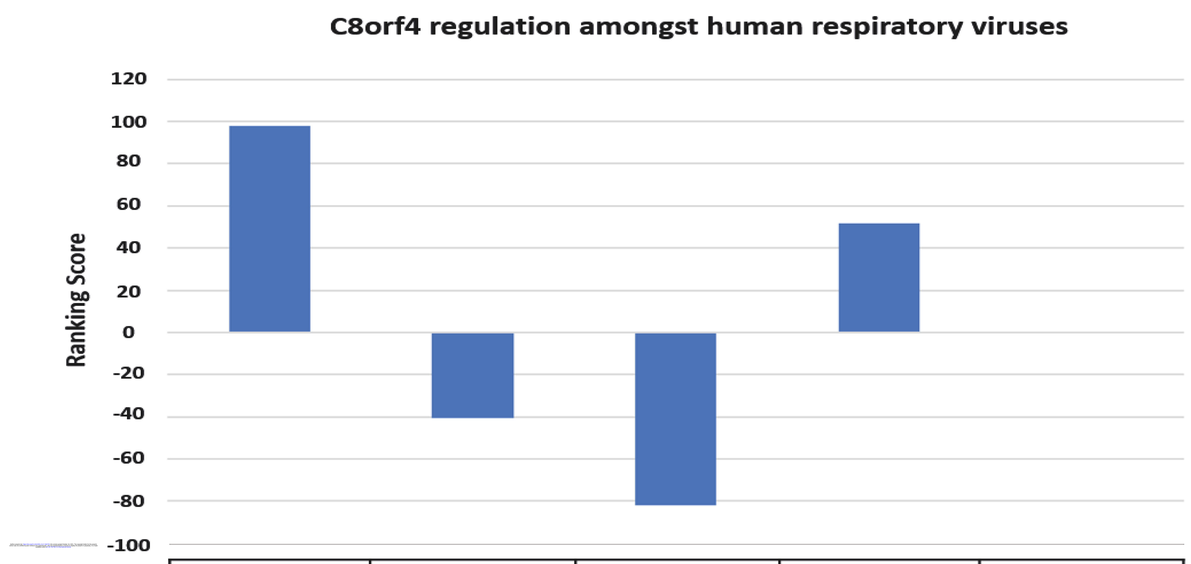
Data on KEGG graph
Rendered by Pathview

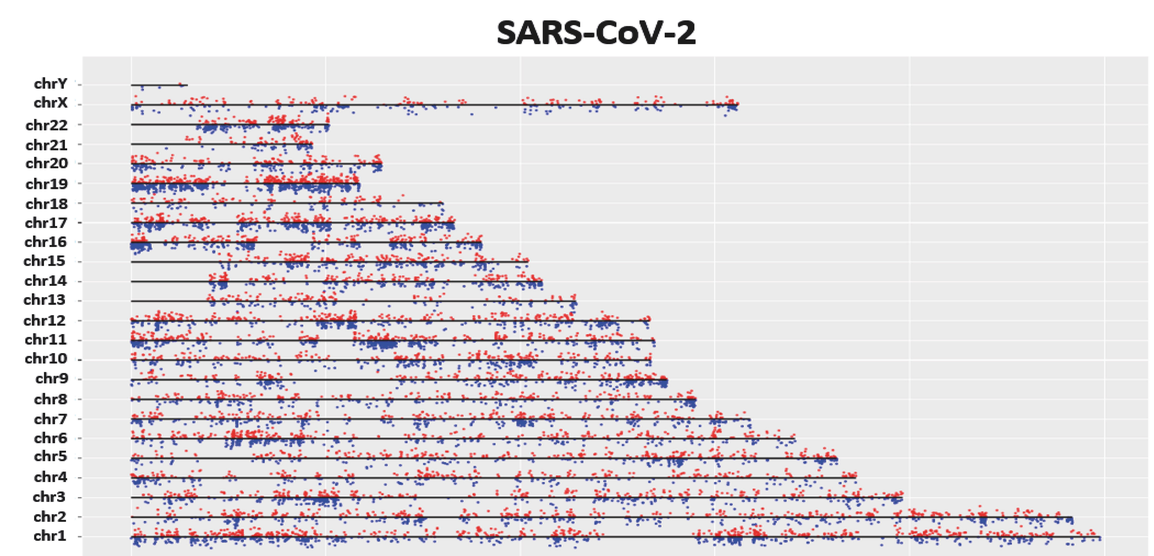
(B)



Data on KEGG graph
Rendered by Pathview

(C)



	SARS-CoV-2	Influenza	SARS-CoV-1	RSV	Rhinovirus
C8orf4	97.85672463	-40.57610242	-82.06065821	51.70222968	0

(D)