Transcriptomics Meta-Analysis Predicts Two Robust Human Biomarkers for Severe Infection with SARS-CoV-2

Jeffrey Clancy¹, Curtis S. Hoffmann¹, Brett E. Pickett¹,*

¹ Department of Microbiology and Molecular Biology; Brigham Young University; Provo, UT, USA.

*Corresponding author: Brett E. Pickett
Corresponding author email: brett_pickett@byu.edu

Keywords: SARS-CoV-2, COVID-19, RNA-sequencing, data mining, biomarkers, RNA, virus, bioinformatics

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Abstract

Defining the human factors associated with severe vs mild SARS-CoV-2 infection has become of increasing interest. Mining large numbers of public gene expression datasets is an effective way to identify genes that contribute to a given phenotype. Combining RNA-sequencing data with the associated clinical metadata describing disease severity can enable earlier identification of patients who are at higher risk of developing severe COVID-19 disease. We consequently identified 358 public RNA-seq human transcriptome samples from the Gene Expression Omnibus database that had disease severity metadata. We then subjected these samples to a robust RNA-seq data processing workflow to quantify gene expression in each patient. This process involved using Salmon to map the reads to the reference transcriptomes, edgeR to calculate significant differential expression levels, and gene ontology enrichment using Camera. We then applied a machine learning algorithm to the read counts data to identify features that best differentiated samples based on COVID-19 severity phenotype. Ultimately, we produced a ranked list of genes based on their Gini importance values that includes GIMAP7 and S1PR2, which are associated with immunity and inflammation (respectively). Our results show that these two genes can potentially predict people with severe COVID-19 at up to ~90% accuracy. We expect that our findings can help contribute to the development of improved prognostics for severe COVID-19.
Introduction

Human infections with severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) has resulted in hundreds of millions of confirmed cases and millions of deaths globally. In addition, countless others have been hospitalized and a subset of the infected population has experienced severe health consequences—particularly those who are elderly, immunocompromised, or have other underlying conditions. The genetic material for this pathogen consists of a monopartite positive-sense single-stranded RNA molecule that is approximately 30 kb in length that contains multiple open reading frames [1]. Since the virus was first detected in late 2019, the scientific community has performed multiple studies to better understand the underlying mechanism(s) of entry and pathogenesis [2–6].

Human pathogenesis studies performed early in the SARS-CoV-2 pandemic showed that the virus induces the interferon response and interleukin-6, as well as other cytokines and chemokines that contribute to COVID-19 [7–9]. Interestingly, multiple studies have shown that the majority of infections are either mild or asymptomatic [10–13]. The large diversity in the human response to infection, combined with large numbers of infections, contributed to strained hospital capacity [14–16]. Various factors contribute to these observed differences in disease severity, and the demand for robust biomarkers associated with COVID-19 disease severity has continually grown. Some studies have identified associations between acute infection and the host response [17,18]. Other studies have evaluated associations between disease severity and aspects of the adaptive immune system [19–23] and quantified viral RNA [24–29]. A recent study has used neural networks to predict patient survival outcomes with high accuracy [30], which can be useful when whole transcriptome data are available. However, to our knowledge, a meta-analysis on transcriptional biomarkers associated with mild versus severe infection has not been previously reported.

The aim of the current study is to perform a meta-analysis of existing human transcriptomics data from collected blood samples to predict transcriptional prognostic markers. Since presence of the virus is possible with existing diagnostics, such markers of disease severity could then contribute to making informed decisions concerning the care of infected patients who seek treatment at the hospital.

Methods

Identification of Relevant Datasets

An established process that combines automated and manual methods was used to identify and analyze the samples from published SARS-CoV-2 human transcriptomics studies with metadata specifying the severity of infection in May, 2021. Samples from patients having either “mild” or “asymptomatic” infections were manually labeled by one reviewer as “mild”, while those recorded as “hospitalized”, “ICU”, or “death” were labeled as “severe”. Records (GSE152418 (PBMC), GSE157103 (whole blood), GSE166424 (whole blood)) in the Gene Expression Omnibus (GEO) database [31–33], within the National Center for Biotechnology Information (NCBI), were queried and manually reviewed for relevance (Figure 1). Specifically, studies were selected based on the
presence of three predefined criteria: 1) the host organism was human, 2) the data were generated as part of a RNA-sequencing experiment, and 3) the study included samples collected during acute SARS-CoV-2 infection together with disease severity metadata. In total, we identified and processed 358 relevant samples across three independent RNA-sequencing studies.

Data Pre-Processing

The fastq files containing the RNA-sequencing data were obtained from the Sequence Read Archive (SRA) at NCBI [34] using the sratools software. These files were divided into “case” and “control” categories based on whether they had “severe” or “mild” disease. The Automated Reproducible MOdular Workflow for Preprocessing and Differential Analysis of RNA-seq Data (ARMOR) was then used to preprocess and analyze the RNA-seq data [35]. Briefly, this ARMOR workflow uses the python-based snakemake workflow language [36] to perform steps including: trimming of sequencing adapters and low-quality regions from the originally-generated RNA-sequencing reads with TrimGalore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), calculate quality control metrics with FastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc/), as well as map and quantify reads to the human GRCh38 transcriptome with Salmon [37]. Differential gene expression was calculated from the read counts with edgeR [38]. Gene ontology terms were calculated from the list of significant gene identifiers produced by edgeR by using the DAVID resource [39].

Machine Learning

The salmon counts for each gene in each GEO sample identifier were compiled into a table. The counts for each sample were then normalized using a z-score transformation. These samples into test and training sets, with 70% of samples assigned to a training set and the remaining 30% of samples assigned to a test set. A random forest classification method (using the R randomforest package) was then trained and used to identify the genes that were most useful in classifying samples as coming from patients with either “severe” or “mild” COVID-19 disease. This random forest algorithm calculated the Gini impurity values for each feature, which were then sorted to rank the importance of genes based on the original training set. This process was repeated for subsets of the best-scoring features to quantify their accuracy, specificity, and sensitivity.

Results

Transcriptomics Meta-Analysis Identifies Significant Genes, Enriched Terms, and Signaling Pathways

We began by identifying publicly available RNA-sequencing data generated from either whole blood samples or peripheral blood mononuclear cells (PBMCs) that had been previously collected from patients infected with SARS-CoV-2 and had associated disease severity metadata. We then assigned these samples to either “high severity” or “low severity”. We processed these RNA-seq files using an automated computational workflow that performed quality control, trimmed reads, mapped them to the human transcriptome, and calculated significant differentially-expressed genes. We then used these genes to identify enriched Gene Ontology (GO) terms.

Overall, we identified 8176 significant differentially expressed genes after applying a multiple hypothesis correction with log₂ fold-change values ranging from -4.2 to 3.78 (Figure 2). We found that the most significant differentially expressed genes included ASPH, C5orf30, DGKH, SLC26A6. We then subjected this list of significant DEGs to Gene Ontology enrichment. This analysis produced 90 significant GO terms including immune response, apoptosis, and I-kappaB
kinase/NF-kappaB signaling (Table 1). It is possible that a subset of these results may contain bias due to the heterogeneity of the samples. However, we expect that the robust statistical analysis of these samples minimized such effects.

We then used the signaling pathway impact analysis algorithm to calculate which intracellular signaling pathways were best represented by the list of DEGs. This analysis identified nine pathways that were significantly affected by severe COVID-19 (Table 2). We observed that five of these significant pathways dealt directly with T-cell receptor (TCR) signaling, while a sixth described a Zap70 immunological synapse. Interestingly, all six of these immune-related pathways were predicted to be inhibited during severe COVID-19.

**Machine Learning Identifies GIMAP7 and S1PR2 as Robust Biomarkers**

Prior to predicting relevant biomarkers, we first wanted to confirm whether the intracellular transcriptional response in blood was strongly associated with the disease severity metadata that was recorded with each record. We consequently constructed a table with all transcripts from each gene represented as columns and the read mapping data represented as rows. We then used this table to generate a receiver-operator characteristic (ROC) curve (Figure 3). In this case, the area under the curve (AUC) represents the percent specificity and sensitivity for the host transcriptomic data to predict disease severity. We calculated the AUC of the curve across all transcripts to be 96.6%, which indicates that the host transcriptional response strongly contributes to disease severity.

We then subjected the same tabular data to a machine learning algorithm to enable us to predict the features (e.g. expressed genes) that were most associated with the disease severity phenotype (Table 3). We ranked our random forest output by descending order of the Mean Decrease in Gini Impurity value, which is a measure of entropy. Transcripts from genes with larger Gini Impurity values represented those that could be used to most accurately predict the recorded disease phenotype.

We next wanted to reduce the selected markers to the smallest number that would provide the best performance. To do so, we generated a ROC curve for the six expressed genes having the highest Gini Impurity values and calculated its AUC to be 94.3%. We repeated this process for only the top two expressed genes (GIMAP7 and S1PR2) and only the top expressed gene (GIMAP7). This analysis quantified the AUC to be 89.8% for the two combined genes and 84.4% for only the top gene. Specifically, the mean and median read counts for each of these two genes were approximately three times higher in the samples with low disease severity than in the samples with high disease severity.

**Figure 1**: Flow diagram to visualize the process used to filter relevant public studies, samples, and files.
PRISMA 2020 flow diagram for new systematic reviews which included searches of databases and registers only

Identification of studies via databases and registers

Identification

Records identified from:
Databases (n = 1001)
Registers (n = 0)

Records removed before screening:
Duplicate records removed (n = 0)
Records marked as ineligible by automation tools (n = 0)
Records removed for other reasons (n = 0)

Screening

Records screened (n = 1001)

Records excluded by automation (n = 945)
Not RNA-sequencing (n = 935)
Not Homo sapiens (n = 10)
No severity metadata (n = 53)

Reports sought for retrieval (n = 358)

Reports not retrieved (n = 0)

Reports assessed for eligibility (n = 358)

Studies included in review (n = 3)
Reports of included studies (n = 358)


For more information, visit: http://www.prisma-statement.org/
<table>
<thead>
<tr>
<th>GO Term</th>
<th>Count</th>
<th>%</th>
<th>P-Value</th>
<th>Corrected P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>protein binding</td>
<td>5518</td>
<td>69.4</td>
<td>3.10E-105</td>
<td>9.70E-102</td>
</tr>
<tr>
<td>RNA binding</td>
<td>731</td>
<td>9.2</td>
<td>2.80E-21</td>
<td>4.30E-18</td>
</tr>
<tr>
<td>identical protein binding</td>
<td>802</td>
<td>10.1</td>
<td>2.20E-14</td>
<td>2.30E-11</td>
</tr>
<tr>
<td>metal ion binding</td>
<td>1188</td>
<td>15</td>
<td>2.90E-13</td>
<td>2.20E-10</td>
</tr>
<tr>
<td>regulation of catalytic activity</td>
<td>215</td>
<td>2.7</td>
<td>7.90E-12</td>
<td>5.10E-08</td>
</tr>
<tr>
<td>apoptotic process</td>
<td>296</td>
<td>3.7</td>
<td>1.10E-11</td>
<td>5.10E-08</td>
</tr>
<tr>
<td>DNA repair</td>
<td>158</td>
<td>2</td>
<td>2.60E-11</td>
<td>8.00E-08</td>
</tr>
<tr>
<td>catalytic activity</td>
<td>89</td>
<td>1.1</td>
<td>6.40E-11</td>
<td>3.90E-08</td>
</tr>
<tr>
<td>regulation of transcription, DNA-templated</td>
<td>456</td>
<td>5.7</td>
<td>7.60E-11</td>
<td>1.80E-07</td>
</tr>
<tr>
<td>positive regulation of I-kappaB kinase/NF-kappaB signaling</td>
<td>115</td>
<td>1.4</td>
<td>2.00E-10</td>
<td>3.60E-07</td>
</tr>
<tr>
<td>ATP binding</td>
<td>710</td>
<td>8.9</td>
<td>5.00E-10</td>
<td>2.60E-07</td>
</tr>
<tr>
<td>protein phosphorylation</td>
<td>255</td>
<td>3.2</td>
<td>1.30E-09</td>
<td>2.00E-06</td>
</tr>
<tr>
<td>Term</td>
<td>Fold</td>
<td>p-value (E)</td>
<td>q-value (E)</td>
<td></td>
</tr>
<tr>
<td>----------------------------------------</td>
<td>------</td>
<td>-------------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>ubiquitin protein ligase binding</td>
<td>168</td>
<td>2.1</td>
<td>5.40E-09</td>
<td></td>
</tr>
<tr>
<td>protein transport</td>
<td>221</td>
<td>2.8</td>
<td>7.50E-09</td>
<td></td>
</tr>
<tr>
<td>protein ubiquitination</td>
<td>244</td>
<td>3.1</td>
<td>9.30E-09</td>
<td></td>
</tr>
<tr>
<td>innate immune response</td>
<td>281</td>
<td>3.5</td>
<td>1.30E-08</td>
<td></td>
</tr>
<tr>
<td>adaptive immune response</td>
<td>217</td>
<td>2.7</td>
<td>1.60E-08</td>
<td></td>
</tr>
<tr>
<td>ubiquitin-protein transferase activity</td>
<td>141</td>
<td>1.8</td>
<td>1.80E-08</td>
<td></td>
</tr>
</tbody>
</table>

*Note: CC-BY 4.0 International license; it is made available under a CC-BY 4.0 International license.*
Table 2: Significant Intracellular Signaling Pathways During Severe COVID-19.

<table>
<thead>
<tr>
<th>Pathway Name</th>
<th>pSize</th>
<th>NDE</th>
<th>pNDE</th>
<th>tA</th>
<th>pPERT</th>
<th>pG</th>
<th>pGFdr</th>
<th>pGFWER</th>
<th>Status</th>
<th>SourceDB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generation of second messenger molecules</td>
<td>30</td>
<td>22</td>
<td>0.0957</td>
<td>0.0027</td>
<td>17.5118</td>
<td>1614</td>
<td></td>
<td></td>
<td>Inhibited</td>
<td>Reactome</td>
</tr>
<tr>
<td>Translocation of ZAP-70 to Immunological synapse</td>
<td>16</td>
<td>11</td>
<td>0.3318</td>
<td>0.0088</td>
<td>8.05626</td>
<td>5319</td>
<td></td>
<td></td>
<td>Inhibited</td>
<td>Reactome</td>
</tr>
<tr>
<td>Degradation of the extracellular matrix</td>
<td>50</td>
<td>32</td>
<td>0.3408</td>
<td>0.0090</td>
<td>20.4183</td>
<td>8026</td>
<td></td>
<td></td>
<td>Activated</td>
<td>Reactome</td>
</tr>
<tr>
<td>Extracellular matrix organization</td>
<td>147</td>
<td>87</td>
<td>0.6251</td>
<td>0.0159</td>
<td>30.6639</td>
<td>7972</td>
<td></td>
<td></td>
<td>Activated</td>
<td>Reactome</td>
</tr>
<tr>
<td>TCR signaling</td>
<td>113</td>
<td>64</td>
<td>0.8034</td>
<td>0.0201</td>
<td>25.4660</td>
<td>7983</td>
<td></td>
<td></td>
<td>Inhibited</td>
<td>Reactome</td>
</tr>
<tr>
<td>Downstream TCR signaling</td>
<td>92</td>
<td>47</td>
<td>0.9689</td>
<td>0.0240</td>
<td>39.9496</td>
<td>465</td>
<td></td>
<td></td>
<td>Inhibited</td>
<td>Reactome</td>
</tr>
<tr>
<td>Role of mef2d in t-cell apoptosis</td>
<td>24</td>
<td>17</td>
<td>0.1945</td>
<td>0.0010</td>
<td>-12.912</td>
<td>83144</td>
<td></td>
<td></td>
<td>Inhibited</td>
<td>BioCarta</td>
</tr>
<tr>
<td>TCR signaling in naive CD8+ T cells</td>
<td>47</td>
<td>37</td>
<td>0.0054</td>
<td>0.0130</td>
<td>61.2685</td>
<td>6962</td>
<td></td>
<td></td>
<td>Inhibited</td>
<td>NCI</td>
</tr>
<tr>
<td>TCR signaling in naive CD4+ T cells</td>
<td>59</td>
<td>45</td>
<td>0.0066</td>
<td>0.0157</td>
<td>34.9932</td>
<td>4227</td>
<td></td>
<td></td>
<td>Inhibited</td>
<td>NCI</td>
</tr>
</tbody>
</table>

pSize: the number of nodes in the pathway.
NDE: number of differentially expressed genes based on unadjusted p-value.
PND: hypergeometric p-value for enriched DEGs in pathway.
tA: total net accumulated perturbation (tA).
pPERT: bootstrap p-value.
pG: unadjusted global probability.
pGFdr: FDR correction of pG < 0.05.
pGFWER: Bonferroni-corrected pG.
Status: predicted effect (Activated/Inhibited) on signaling pathway based on the direction of the tA value.
Table 3: List of top 10 expressed genes that best predicted SARS-CoV-2 disease severity in human blood cells.

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>DEG Log₂FC*</th>
<th>DEG FDR**</th>
<th>Gini Impurity</th>
<th>Severity Level</th>
<th>Mean (Read Counts)</th>
<th>Standard Deviation (Read Counts)</th>
<th>Median (Read Counts)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GIMAP7</td>
<td>-1.16</td>
<td>9.94E-23</td>
<td>0.44</td>
<td>High</td>
<td>1150.79</td>
<td>840.69</td>
<td>932</td>
</tr>
<tr>
<td>S1PR2</td>
<td>-1.27</td>
<td>1.3E-17</td>
<td>0.41</td>
<td>High</td>
<td>44.81</td>
<td>67.55</td>
<td>36</td>
</tr>
<tr>
<td>PRR5L</td>
<td>-1.01</td>
<td>2.6E-17</td>
<td>0.4</td>
<td>High</td>
<td>148.06</td>
<td>111.09</td>
<td>112</td>
</tr>
<tr>
<td>RABGAP1L</td>
<td>-0.745</td>
<td>8.64E-17</td>
<td>0.38</td>
<td>High</td>
<td>834.23</td>
<td>247.90</td>
<td>816.5</td>
</tr>
<tr>
<td>TRERF1</td>
<td>-0.642</td>
<td>5.77E-16</td>
<td>0.35</td>
<td>High</td>
<td>246.87</td>
<td>127.21</td>
<td>222.5</td>
</tr>
<tr>
<td>GPR174</td>
<td>-1.32</td>
<td>2.49E-23</td>
<td>0.33</td>
<td>High</td>
<td>109.62</td>
<td>93.16</td>
<td>81</td>
</tr>
<tr>
<td>CRTAM</td>
<td>-1.01</td>
<td>4.26E-19</td>
<td>0.32</td>
<td>High</td>
<td>52.33</td>
<td>38.58</td>
<td>43.5</td>
</tr>
<tr>
<td>GPR68</td>
<td>-1.35</td>
<td>9.96E-18</td>
<td>0.31</td>
<td>High</td>
<td>30.68</td>
<td>29.67</td>
<td>22.5</td>
</tr>
<tr>
<td>CD2</td>
<td>-1.28</td>
<td>1.37E-18</td>
<td>0.3</td>
<td>High</td>
<td>567.39</td>
<td>564.07</td>
<td>449.5</td>
</tr>
</tbody>
</table>

The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.
Figure 2. Volcano plot of all differentially expressed genes in severe vs. mild human infection with SARS-CoV-2. Genes that are up or down regulated from blood samples collected from patients having severe symptoms or mild symptoms during infection with SARS-CoV-2. Genes showing statistically significant up-regulation (blue), down-regulated (red), or no significant change (green). X-axis shows the log$_2$ fold-change values while the y-axis displays false-discovery rate-adjusted p-values to account for multiple hypothesis testing.
Figure 3. Receiver-operator characteristic (ROC) curve constructed from all expressed genes in severe vs. mild human infection with SARS-CoV-2. Constructing a ROC curve from all RNA-sequencing read quantification values achieved an area-under-the-curve (AUC) value of greater than 96%.

Discussion

Previous work has shown that combining multiple datasets in meta-analyses can augment the signal-to-noise ratio in order to gain new biological and mechanistic insight(s) [9,40–43]. As such, the goal of our study was to predict prognostic markers of SARS-CoV-2 disease severity through a transcriptomics meta-analysis of human blood samples. We found statistically significant differentially expressed genes, enriched Gene Ontology terms, modulated signaling pathways, and a ranked list of biomarkers that could potentially be combined to predict which patients are at risk of severe disease.

At least one prior study has generated single-cell RNA-sequencing data to better understand the host response to SARS-CoV-2 infection [44]. Their findings showed adaptive immune components play a role in disease severity. Interestingly, our data confirm results from some prior experiments that show certain aspects of the T-cell response may be downregulated during SARS-CoV-2 infection [45–47]. A modified distribution of the ZAP70 kinase on the plasma membrane of T-cells contributes to the signal transduction and amplification of the TCR [48]. Interestingly, CD3/ZAP70 protein has been shown to interact with TREM-2 in the T-cells of
patients infected with SARS-CoV-2 [49]. The ASPH and C5orf30 genes, which were identified as significant DEGs in our study, have been identified previously as markers of severe infection [50].

These prior studies support our finding that GIMAP7 could be important in severity given its presence on the surface of T-lymphocytes [51]. In the time since GIMAP7 was initially identified [52], it has subsequently been shown to be a potential marker for various cancer types, which further supports its immune-related role [53–56]. Although the underlying mechanism for this protein is yet to be characterized, we are not surprised by its role in other immune-related diseases or in SARS-CoV-2 disease severity. Information about the S1PR2 gene is more sparse, although it appears to have a role in inflammation and other immune-related functions [57–61]. Specifically, S1PR2 on T-cells recruits lymphocytes to damaged tissues and may contribute to the recirculation of cells in the adaptive immune system to the lymphatic system [62].

The GIMAP7 gene, together with a subset of the enriched GO terms, had been found in an earlier study on transcriptional biomarkers for SARS-CoV-2 [63]; however, it was not highly ranked. This is logical since DEGs are often identified as biomarkers. In addition, as the sample size increases among a more diverse patient population, we expect that the genes most capable of differentiating disease severity should become more accurate.

It is important to note that these samples were taken while the early variants were circulating. As such, further testing would be required to confirm whether these biomarkers are still consistent predictors of infection severity in patients infected with more recent variants. We anticipate that qRT-PCR (quantitative Reverse Transcriptase Polymerase Chain reaction), and possibly flow cytometry could be used to quantify these biomarkers in relevant samples. Additional experiments are needed to confirm whether these findings can be replicated in samples across different age and/or risk groups.

These results present a potential predictor for the severity levels of patients infected with SARS-CoV-2. Considering the diversity of reactions that patients have to the virus, incorporating these biomarkers as additional data points to assess patient risk of severe disease could be pivotal in augmenting both personal and public health decision making processes [64–66]. This is especially relevant when resources may be limited and priority must be given to those with the greatest risk of severe infection. These results in particular are useful because they illustrate both up and down regulation association that best differentiate each patient in a population, rather than just identifying genes with statistically significant changes across the populations being compared. This approach allows us to detect the directionality of biomarkers such as GIMAP7 and S1PR2 that were not as highly ranked by edgeR.

In conclusion, the findings from this study could contribute to ongoing efforts relating to triaging patients when hospitals are approaching their capacity limit. We envision that creating an assay to quantify the presence of a subset of the features identified in this study could be useful in identifying patients who are at higher risk of developing severe disease. The results of such a prognostic assay could contribute to triage efforts and to treatment decisions being made in the clinic.

**Acknowledgements**

We are grateful to the BYU Office of Research Computing for providing the computational resources needed to complete this study. We also thank the original clinicians, patients, and others who provided the RNA-sequencing data.

**References**


16. Finelli L, Gupta V, Petigara T, Yu K, Bauer KA, Puzniak LA. Mortality Among US Patients


PRISMA 2020 flow diagram for new systematic reviews which included searches of databases and registers only

**Identification of studies via databases and registers**

Records identified from:
- Databases (n = 1001)
- Registers (n = 0)

Records removed before screening:
- Duplicate records removed (n = 0)
- Records marked as ineligible by automation tools (n = 0)
- Records removed for other reasons (n = 0)

Records screened (n = 1001)

Records excluded by automation:
- (n = 945)
  - Not RNA-sequencing (n = 935)
  - Not Homo sapiens (n = 10)
  - No severity metadata (n = 53)

Reports sought for retrieval (n = 358)

Reports not retrieved (n = 0)

Reports assessed for eligibility (n = 358)

Reports excluded: (n = 0)

Studies included in review (n = 3)

Reports of included studies (n = 358)

---


For more information, visit: [http://www.prisma-statement.org/](http://www.prisma-statement.org/)

---

**Figure 1**