bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.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.

Host transcriptional responses and SARS-CoV-2 isolates from the nasopharyngeal samples of Bangladeshi COVID-19 patients

Abul Bashar Mir Md. Khademul Islam[1*#], Md. Abdullah-Al-Kamran Khan[2#], Rasel Ahmed[3], Md. Sabbir Hossain[3], Shah Md. Tamim Kabir[3], Md. Shahidul Islam[3], A.M.A.M. Zonaed Siddiki[4]

[1]Department of Genetic Engineering & Biotechnology, University of Dhaka, Dhaka, Bangladesh.

[2]Department of Mathematics and Natural Sciences, BRAC University, Dhaka, Bangladesh.

[3]Basic and Applied Research on Jute Project, Bangladesh Jute Research Institute, Dhaka, Bangladesh.

[4]Department of Pathology and Parasitology, Chittagong Veterinary and Animal Sciences University (CVASU), Khulshi, Chittagong, Bangladesh.

[#] **These authors contributed equally.**

* Correspondence:

Dr. Abul Bashar Mir Md. Khademul Islam
Associate Professor
Department of Genetic Engineering and Biotechnology
University of Dhaka
Dhaka 1000, Bangladesh.
Email: khademul@du.ac.bd

Keywords

Host transcriptional response, COVID-19, SARS-CoV-2, Genome variations, Immune response, Integrins

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.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

As the COVID-19 pandemic progresses, fatality and cases of new infections are also increasing at an alarming rate. SARS-CoV-2 follows a highly variable course and it is becoming more evident that individual's immune system has a decisive influence on the progression of the disease. However, the detailed underlying molecular mechanisms of the SARS-CoV-2 mediate disease pathogenesis are largely unknown. Only a few host transcriptional responses in COVID-19 have been reported so far from the Western world, but no such data has been generated from the South-Asian region yet to correlate the conjectured lower fatality around this part of the globe. In this context, we aimed to perform the transcriptomic profiling of the COVID-19 patients from Bangladesh along with the reporting of the SARS-CoV-2 isolates from these patients. Moreover, we performed a comparative analysis to demonstrate how differently the various SARS-CoV-2 infection systems are responding to the viral pathogen. We detected a unique missense mutation at 10329 position of ORF1ab gene, annotated to 3C like proteinase, which is found in 75% of our analyzed isolates; but is very rare globally. Upon the functional enrichment analyses of differentially modulated genes, we detected a similar host induced response reported earlier; this response was mainly mediated by the innate immune system, interferon stimulation, and upregulated cytokine expression etc. in the Bangladeshi patients. Surprisingly, we did not perceive the induction of apoptotic signaling, phagosome formation, antigen presentation and production, hypoxia response within these nasopharyngeal samples. Furthermore, while comparing with the other SARS-CoV-2 infection systems, we spotted that lung cells trigger the more versatile immune and cytokine signaling which was several folds higher compared to our reported nasopharyngeal samples. We also observed that lung cells did not express *ACE2* in a very high amount as suspected, however, the nasopharyngeal cells are found overexpressing *ACE2*. But the amount of *DPP4* expression within the nasal samples was significantly lower compared to the other cell types. Surprisingly, we observed that lung cells express a very high amount of integrins compared to the nasopharyngeal samples, which might suggest the putative reasons for an increased amount of viral infections in the lungs. From the network analysis, we got clues on the probable viral modulation for the overexpression of these integrins. Our data will provide valuable insights in developing potential studies to elucidate the roles of ethnicity effect on the viral pathogenesis, and incorporation of further data will enrich the search of an effective therapeutics.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. 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.

Introduction

Since the declaration of COVID-19 pandemic on 11 March, this Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) mediated infection has spread ~213 countries and territories [1]. Approximately, 15 million individuals across the globe have fallen victim to this virus and the number is constantly increasing at an alarming rate, as of the writing of this manuscript [1]. Though the initial fatality percentage was as low as 3.5%, currently this value lies around ~6.66% [1] and it might be increased because of the withdrawal of earlier preventing measures taken throughout the world. Coronaviruses are not new to human civilization, as these viruses caused several earlier outbreaks during the past two decades. However, none of the earlier outbreaks spread as widely as the current ongoing pandemic. As the pandemic progresses, more researches on the molecular pathobiology of the COVID-19 are being rapidly carried out to search for effective therapeutic intervention.

Coronaviruses possess single-stranded RNA (positive sense) genomes lengthening approximately 30Kb [2]. Amongst the coronaviruses, SARS-CoV-2 is a member of the betacoronaviruses having a ~29.9Kb genome which contains 11 functional genes [3]. Though SARS-CoV-2 shows similar clinical characteristics as Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) and Middle East Respiratory Syndrome-related Coronavirus (MERS-CoV) viruses, it has only ~79% and ~50% genome sequence similarities with these viruses, respectively; whereas, the genome sequence of SARS-CoV-2 is ~90% identical to that of bat derived SARS-like coronavirus [4]. Moreover, several key genomic variances between SARS-CoV-2 and SARS-CoV such as- 380 different amino acid substitutions, ORF8a deletion, ORF8b elongation, and ORF3b truncation were also reported [2].

The clinical characteristics of the COVID-19 range from mild fever to severe lung injury [5]. Some of the commonly observed mild COVID-19 symptoms are- fever, cough, and fatigue; however, complications such as- myalgia, shortness of breath, headache, diarrhea, and sore throat were also reported [6]. Furthermore, severely affected patients had exhibited respiratory complications like- moderate to severe pneumonia, acute respiratory distress syndrome (ARDS), sepsis, acute lung injury (ALI), and multiple organ dysfunction (MOD) [7]. Primarily, the lungs of the COVID-19 patients are affected [8]; however, failures of other functional systems, namely- cardiovascular system, nervous system etc. were also reported [9, 10].

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. 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.

Several features of the SARS-CoV-2 infection made it more complicated for effective clinical management. From the earlier studies, the incubation period of SARS-CoV-2 was reported to be around 4-5 days, however, some recent studies suggested a prolonged incubation period of 8-27 days [11]. Additionally, several cases of viral latency within the host [12], and the recurrent presence of SARS-CoV-2 in clinically recovered patients were also recorded [13, 14]. However, the detailed molecular mechanism behind these phenomena is still elusive.

In COVID-19, an increased level infection associated pro-inflammatory cytokines were recorded [15], which thereby supports the term “Cytokine storm”, that was frequently used to describe the SARS-CoV and MERS-CoV disease pathobiology [16]. This phenomenon causes the hyperactivation and recruitment of the inflammatory cells within the lungs and results in the acute lung injury of the infected patients [17]. However, this illustrates one putative molecular mechanism of COVID-19, there are many other immune regulators and host genetic/epigenetic factors which can also play significant contribution towards the disease manifestation [18, 19]. This multifaceted regulation was also reported previously for other different coronavirus infections [20]. Host-pathogen interactions in different coronavirus infections can function as a double-edged sword, as these could be beneficial not only to the hosts but also the viruses [20]. Similar host-virus tug-of-war can also occur in COVID-19 which might be contributing towards the overcomplicated disease outcomes [21].

Collectively, more than 1.7 million (almost 9% of the total infections around the globe) people have been diagnosed with COVID-19 in the South-Asian region and the number is still increasing devastatingly [1]. Recently, it has been speculated that South-Asian people might be possessing a genomic region acting as the risk factor for COVID-19 [22]. Moreover, another study suggested some genomic variations in several Indian SARS-CoV-2 isolates that might be involved in the COVID-19 pathogenesis in Indian patients [23]. However, any data suggesting the COVID-19 patients’ transcriptomic responses from this part of the globe are yet to be reported.

Several previously conducted studies reported the host transcriptional responses in SARS-CoV-2 infections using patient samples, animal models, and cell lines to explain the pathobiology of COVID-19 [24-26]. However, a detailed comparison of the host transcriptional responses between these different infection models as well as the different sites of the respiratory system is still lacking; but it might provide valuable insights on the COVID-19 pathogenesis and disease severity. In this present study, we sought to discuss the

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. 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.

host transcriptional responses observed in COVID-19 patients in Bangladesh. Additionally, we reported the genome variations observed in the four SARS-CoV-2 isolates obtained from these patients. Finally, we illuminated the differences in host transcriptional responses in different COVID-19 infection models and further pursued to discover the putative effects of these altered responses.

Materials and Methods

Sample collection and virus detection by Real-time reverse transcription-quantitative PCR (RT-qPCR)

The nasopharyngeal swab samples were collected from patients suspicious of COVID-19 and placed in sample collection vial containing normal saline. Collected samples were preserved at -20°C until further use for RNA extraction and RT-qPCR assay. The RT-qPCR was performed for ORF1ab and N genes of SARS-CoV-2 using Novel Coronavirus (2019-nCoV) Nucleic Acid Diagnostic Kit (PCR-Fluorescence Probing) of Sansure Biotech Inc. according to the manufacturer's instructions. RNA was extracted from a 20 µL swab sample through lysis with sample release reagent provided by the kit and then directly used for RT-qPCR. Thermal cycling was performed at 50 °C for 30 min for reverse transcription, followed by 95 °C for 1 min and then 45 cycles of 95 °C for 15 s, 60 °C for 30 s on an Analytik-Jena qTOWER instrument (Analytik Jena, Germany).

RNA sequencing

Total RNA was extracted from nasopharyngeal swab samples (labeled as S2, S3, S4, S9) collected from SARS-COV-2 infected COVID-19 patients using TRIzol (Invitrogen) reagent following the manufacturer's protocol. RNA-seq libraries were prepared from total RNA using TruSeq Stranded Total RNA Library Prep kit (Illumina) according to the manufacturer's instructions where the first-strand cDNA was synthesized using SuperScript II Reverse Transcriptase (Thermo Fisher) and random primers. Paired-end (150 bp reads) sequencing of the RNA library was performed on the Illumina NextSeq 500 platform.

Data processing and identification of the viral agent

Firstly, the sequencing reads were adapter and quality trimmed using the Trimmomatic program [27]. The remaining reads were mapped against the SARS-CoV-2 reference

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. 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.

sequence (NC_045512.2) using Bowtie 2 [28]. Then the mapped reads were assembled de novo using Megahit (v.1.1.3) [29].

Mapping of the RNA-seq reads onto SARS-CoV-2 reference genome

We mapped the normalized (by count per million mapped reads-CPM) RNA-seq reads onto the SARS-CoV-2 genome track of the UCSC genome browser [30] using the “bamCoverage” feature of deepTools2 suite [31].

Identification of SARS-CoV-2 genome variations and variation annotation

We identified the variations within our sequenced SARS-CoV-2 genome using the “Variation Identification” (https://bigd.big.ac.cn/ncov/online/tool/variation) tool of “2019 Novel Coronavirus Resource (2019nCoVR)” portal of China National Center for Bioinformation [32]. We then annotated the variations of the isolated SARS-CoV-2 isolates using the “Variation Annotation” (https://bigd.big.ac.cn/ncov/online/tool/annotation) tool from the same portal [32]. We also gathered the global frequency of every identified variation using this same information portal [32]. Different representations showing the information regarding the variations were produced using the Microsoft Excel program [33]. The impacts of the variations were further characterized utilizing the Ensembl Variant Effect Predictor (VEP) tool [34].

Analysis of RNA-seq expression data

We analyzed both our RNA-seq and some publicly available RNA-seq data for COVID-19 host transcriptional profile analysis. Publicly available Illumina sequenced RNA-seq raw FastQ reads were extracted from the GEO database (accessions of the data used can be found in supplementary file 1) [35]. We have checked the raw sequence quality using FastQC program (v0.11.9) [36] and found that the "Per base sequence quality", and "Per sequence quality scores" were high over the threshold for all sequences (Supplementary file 2). The mapping of reads was done with TopHat (tophat v2.1.1 with Bowtie v2.4.1) [37]. Short reads were uniquely aligned allowing at best two mismatches to the human reference genome from (GRCh38) as downloaded from the UCSC database [38]. Sequence matched exactly more than one place with equally quality were discarded to avoid bias [39]. The reads that were not mapped to the genome were utilized to map against the transcriptome (junctions mapping). Ensembl gene model [40] (version 99, as extracted from UCSC) was used for this process. After mapping, we used the SubRead package featureCount (v2.21) [41] to calculate absolute

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. 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.

read abundance (read count, rc) for each transcript/gene associated to the Ensembl genes. For differential expression (DE) analysis we used DESeq2 (v1.26.0) with R (v3.6.2; 2019-07-05) [42] that uses a model based on the negative binomial distribution. To avoid false positive, we considered only those transcripts where at least 10 reads are annotated in at least one of the samples used in this study and also applied a minimum Log2 fold change of 0.5 for to be differentially apart from adjusted p-value cut-off of ≤ 0.05 by FDR. To assess the fidelity of the RNA-seq data used in this study and normalization method applied here, we checked the normalized Log2 expression data quality using R/Bioconductor package "arrayQualityMetrics (v3.44.0)" [43]. From these analyses, no outlier was detected in our data by "Distance between arrays", "Boxplots", and "MA plots" methods and replicate samples are clustered together (data not shown). We also performed a multifactorial differential gene expression analysis using the edgeR tool [44] following the experimental design- (Sample A/control for sample A)/(Sample B/control for sample B).

Construction of phylogenetic tree

We constructed a Neighbour-Joining phylogenetic tree with all available 145 SARS-CoV-2 genomes of Bangladeshi isolates (retrieved on 6th May from GISAID [45]). Firstly, the genome sequences were aligned using MAFFT [46] tool using the auto-configuration. Then we used MEGA X [47] for constructing the phylogenetic tree utilizing 500 bootstrapping with substitution model/method: maximum composite likelihood, uniform rates of variation among sites, the partial deletion of gaps/missing data and site coverage cutoff 95%.

Functional enrichment analysis

We utilized Gitools (v1.8.4) for enrichment analysis and heatmap generation [48]. We have utilized the Gene Ontology Biological Processes (GOBP) [49], Bioplanet pathways [50], KEGG pathway [51], and Reactome pathway [52] modules for the overrepresentation analysis. Resulting p-values were adjusted for multiple testing using the Benjamin and Hochberg's method of False Discovery Rate (FDR) [53].

Retrieval of the host proteins that interact with SARS-CoV-2

We have obtained the list of human proteins that form high confidence interactions with SARS-CoV-2 proteins from conducted previously study [21] and processed their provided protein names into the associated HGNC official gene symbols.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. 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.

Construction of biological networks

Construction, visualization, and analysis of biological networks with differentially expressed genes, their associated transcription factors, and interacting viral proteins were executed in the Cytoscape software (v3.8.0) [54]. We used the STRING [55] database to extract the highest confidences (0.9) edges only for the protein-protein interactions to reduce any false positive connection.

Results

Our sequenced SARS-CoV-2 isolates showed a divergent variation pattern compared to the other Bangladeshi isolates

We sought to find out the genome variations within the four SARS-CoV-2 isolates we sequenced and pursued the deviation of these genomes compared to the other isolates in Bangladesh. To accomplish these goals, we first identified and annotated the genome variations observed within our sequenced isolates. Then we produced informative statistics from these observed variations and compared the prevalence of those with the other isolates of Bangladesh and the rest of the world.

We mapped the RNA-seq reads of each of the samples and checked their distribution athwart the entire reference genome of SARS-CoV-2 (Figure 1A). High coverages and read evidence were observed for all the isolates across the whole genome of the SARS-CoV-2 (Figure 1A). This suggests that the sequenced genomes of these isolates are of high coverage and no such region is observed without the mapped reads.

We detected sixty different types of variations within these four analyzed SARS-CoV-2 isolates (Table 1). All the four different types of sequence variations were spotted in these isolates, however, single nucleotide polymorphisms (SNPs) were most prominent (Figure 1B). Among these variations, twelve variations were found in more than one isolate, whereas rest forty-eight variation occurred in only one isolate (Table 1, Figure 1C). Among the isolates, isolate S3 contained the lowest number of variations, whereas isolate S4 has the highest number of variations (Figure 1D). Most of these variations are intergenic variants that occurred either in the 5’-UTR or 3’-UTR regions (Figure 1C-D); whereas, globally most variations have occurred in the ORF1ab gene. Surprisingly, out of all these variations, we found only one downstream gene variation on the 3’-UTRs of all the four isolates; this

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. 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.

variation can potentially impact the regulation of the ORF10 gene (Figure 1D). Most of the nucleic acid mutations were located on the 3'-UTR of the isolates, whereas the ORF1ab gene contained most of the amino acid mutations (Figure 1E).

No highly severe mutation was identified amongst these variations, but we found nine moderately impacting, seven low impacting, and forty-seven modifier variations within these isolates (Figure 1F, Supplementary file 3). As of 8th July, thirty-eight out of the sixty variations within our sequenced isolates were completely absent in all other Bangladeshi SARS-CoV-2 isolates (Table 1). Strikingly, we observed that variation 10329: A>G is present within three of our sequenced isolates, only one other Bangladeshi and one other USA isolate contain this variation (Figure 1G). This variation is located within the 3C-like protease of SARS-CoV-2. Previously, the potential implication of the mutations of this protein was reported to alter its overall structure and functionality [56-58] in SARS-CoV. Also, few of our reported variations like 25505: A>T and 29392: G>T are not highly prevalent globally (Figure 1G).

Exploring the Nextstrain portal [59], we noticed that our analyzed SARS-CoV-2 sequences are closely placed to the Saudi-Arabian isolates (Supplementary Figure 1); although, most of the other Bangladeshi isolates were placed in the major European clusters (data not shown). Furthermore, these isolates analyzed in this study is distinctly placed in our constructed Neighbor-Joining phylogenetic tree (Figure 2), this also supports the differences between these isolates and other Bangladeshi SARS-CoV-2 isolates which might have been originated from the European nations.

Stimulated antiviral immune responses are detected in the nasopharyngeal samples of Bangladeshi COVID-19 patients

Though initial researches suggested the potential implication of viral variations on the COVID-19 disease severity, one recent study indicated otherwise; Several host factors such as- abnormal immune responses, cytokine signaling etc. might be influencing the overall disease outcomes more prominently compared to the viral mutations [60]. Moreover, several data surmised that ethnicity might be a pivotal risk factor of being susceptible to COVID-19 [61].

In this context, we explored the transcriptome data obtained from the nasopharyngeal samples from Bangladeshi COVID-19 patients to find out how these patients were

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. 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.

responding against the invading SARS-CoV-2. We compared the RNA-seq data of these patients with some random normal individuals' nasopharyngeal RNA-seq data to find out the differentially expressed genes within our analyzed samples. We observed a roughly constant standard deviation for the normalized reads suggesting a lesser amount of variation occurred during the normalization (Figure 3A). Furthermore, we performed sample clustering to assess the quality of our generated normalized RNA-seq data. No anomalies were observed in the sample to sample distance matrix (Figure 3B) and principal component analysis (PCA) (Figure 3C) while comparing our samples with the used healthy individuals' data. Moreover, the larger differences observed in the PCA plot (Figure 3C) and clustered heatmap of the count matrix with the top 50 significant genes (Figure 3D) suggest a significant transcriptomic response difference between our infected patients' data and the normal individuals' data. Likewise, the sample to sample distance plot suggested the similarities of samples of similar nature; the infected and healthy samples were clustered into two distinct groups (Figure 3B).

Sungnak et al. described the significance of several viral entry associated host proteins in SARS-CoV-2 pathogenesis, namely- ACE2, TMPRSS2, BSG, CTSL, DPP4 [62]. We also investigated the expression of the associated transcripts of these proteins within our patients' samples. We spotted that both the healthy and infected samples have expressed these genes except the *DPP4* gene (Figure 3E).

We identified 1,614 differentially expressed genes within our reported four SARS-CoV-2 infected nasopharyngeal samples; among these differentially expressed genes, 558 genes were upregulated, and 1056 genes were downregulated (Supplementary file 4). Then we sought to discover in which biological functions/pathways these deregulated genes might be involved. To achieve this, we performed functional enrichment analyses with the observed deregulated genes using different ontology and pathway modules.

Several GOBP terms related to antiviral immune responses such as- viral process, defense response to virus, innate immune response, inflammatory response, negative regulation of viral transcription, negative regulation of viral genome replication etc. were observed enriched for the upregulated genes (Figure 3F, Supplementary Figure 2). Surprisingly, several other important antiviral defense related functions such as- apoptosis, antigen processing, and presentation etc. were found enriched for downregulated genes (Figure 3F). Similarly, this pattern was also observed for the functional enrichment using KEGG and Bioplanet pathways

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. 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.

modules. Upregulated genes are observed involved in signaling pathways such as- innate immune system, antiviral mechanism by interferon-stimulated genes, interleukin-2 signaling, interferon-gamma signaling, interferon alpha-beta signaling, antiviral mechanism by interferon stimulated genes, IL-17 signaling pathway, Toll-like receptor signaling pathway, RIG-I like receptor signaling pathway, and MAPK signaling pathway, etc. (Figure 3G-H, Supplementary Figure 2). Strikingly, several important antiviral signaling pathways such as- antigen processing and presentation, apoptosis, HIF-1 signaling pathway, Natural killer cell mediated cytotoxicity, phagosome, PI3K-Akt signaling pathway, Interleukin-6 regulation of target genes, and Interleukin-10 inflammatory signaling pathway were enriched for the downregulated genes (Figure 3G-H). This unusual observation made us curious to search for a similar pattern of deregulated host responses in several other COVID-19 disease models.

Host responses observed in nasopharyngeal samples are significantly different compared to the other SARS-CoV-2 infections models

We sought to compare the host responses of our analyzed samples with several other different SARS-CoV-2 infection models (two different lung biopsy samples from COVID-19 patients and two different cell lines). We performed functional enrichment analyses using differentially expressed genes from four other SARS-CoV-2 infection systems and compared the enriched terms of our samples with these four other samples. Moreover, how the host responds differently in different tissue types were also evaluated. To achieve these goals, we identified the differentially expressed genes across these different samples and systematically compared the enrichment results of those deregulated genes.

Using the similar parameterization of the differential gene expression analyses, we identified 6714 genes in lung cells (GSE147507), 232 genes in lung cells (GSE150316), 143 genes in NHBE cells (GSE147507), and 5637 genes in Calu-3 cells (GSE148729) as differentially expressed compared to their respective healthy controls (Supplementary file 5). Significant proportions of the deregulated genes detected in our nasopharyngeal samples are also found deregulated in lung (GSE147507) and Calu-3 cells (GSE148729) samples (Figure 4A), while a small number of our samples' deregulated genes were also observed deregulated in rest of the two samples used (Figure 4A).

Enrichment analysis using these deregulated genes suggested the host response differences among the different infection systems used (Figure 4B-D). Upon the analysis, only a few GOBP terms were found enriched for both our samples, lung (GSE147507), and Calu-3 cells

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. 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.

(GSE148729) samples, such as- viral process, immune response, innate immune response, defense response to virus, interferon signaling etc. (Figure 4B). However, genes in many important antiviral immune response related functions were not significantly enriched for our samples but were enriched for the lung (GSE147507), and Calu-3 cells (GSE148729) samples; these processes are- autophagy, apoptotic signaling pathway, interleukin-6 mediated signaling pathway, interleukin-12 mediated signaling pathway, cytokine-mediated signaling pathway, inflammatory response etc. (Figure 4B). Moreover, processes like- response to hypoxia, response to vitamin-D, and lung development were also not enriched for the deregulated genes of our nasal samples (Figure 4B).

We noticed several commonly enriched important immune signaling pathways for most of the samples used for the comparison (Figure 4C-D), such as adaptive immune system, innate immune system, interferon signaling, apoptosis, Toll-like receptor signaling pathway regulation, antigen processing and presentation, integrin signaling pathway, RIG-I like receptor signaling pathway, phagosomes etc. (Figure 4C-D, Supplementary Figure 3); however, pathways like JAK-STAT signaling pathway, Natural killer cell mediated cytotoxicity, NF-κB signaling pathway, asthma, PI3K-Akt pathway, cellular response to hypoxia, inflammasomes, and inflammatory response pathway etc. (Figure 4C-D, Supplementary Figure 3) were not enriched for the deregulated genes of our nasopharyngeal samples. These results suggest that host responses observed in nasopharyngeal samples have a different host response compared to the other infection systems. Therefore, to unveil the mystery behind this observation, we further analyzed these data to compare the gene expression patterns in different specific functionalities.

Significant gene expression differences were spotted between the nasopharyngeal samples and lung biopsy samples

We compared the normalized read counts of each sample without integrating the respective controls to shed insights on the differences in gene expression patterns between the individual samples and tissues. A constant standard deviation was observed for the normalized read counts of the infected samples (Figure 5A) indicating the acceptability of the normalized reads for analysis. From the sample to sample distance clustering, principal component analysis, and clustered heatmap of the count matrix with top 50 genes, we observed that gene expression profiles of our nasopharyngeal samples are more relevant to that of lung samples; whereas, high level of variance was observed between the gene expression counts of the cell

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. 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.

lines and primary nasopharyngeal samples (Figure 5B-D). Furthermore, we had a similar observation from the clustered normalized read counts of the samples based on Pearson's correlation distance with all genes that vary across samples (Figure 5E). We then narrowed down our searches to the sample level gene expression profiles of several COVID-19 related important biological functions within these samples (Figure 6), to understand the gene expression similarities and dissimilarities among these infections systems, specially comparing nasal and lung tissues.

Genes related to integrins and integrin signaling pathway are highly expressed in lung samples compared to the nasopharyngeal samples

Several previous reports suggested an important aspect of integrins in SARS-CoV-2 pathogenesis, therefore, we sought to find out the expression profiles of integrin related genes in different COVID-19 infection models at sample level [63, 64]. RGD motif of the spike protein of SARS-CoV-2 can bind the integrins and this motif is placed near to the ACE2-receptor binding motif [63]. Moreover, evidence of integrin domain binding was also reported for SARS-CoV [65]. Therefore, we sought to discover the expression profiles of the integrin related genes in different SARS-CoV-2 infection models. To accomplish this, we filtered out the integrin and integrin signaling related genes (Supplementary file 6) within the terms of the GOBP, KEGG pathway, and Bioplanet pathway modules that we used for enrichment analysis. Intriguingly, we observed that the genes related to integrins and integrin signaling were highly expressed in analyzed lung samples, and the lowest number of these genes were expressed in the nasopharyngeal samples (Figure 6A-B, Supplementary Figure 4A). Based on these observations, we can assume that overexpression of integrins and integrin signaling related genes in the lungs might provide the virus a competitive edge in invading the lung cells more efficiently compared to the cells of the nasopharynx and respiratory tracts.

Cytokine and inflammatory signaling genes are overexpressed in lung samples

Aberrant cytokine stimulation and inflammatory responses are thought to be the major contributor to pathogenic lung damages in severely affected COVID-19 patients [66, 67]. We wanted to find out whether the genes related to cytokine signaling and inflammation have differential expression profiles in lung cells compared to the other infection systems. We

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. 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.

extracted and compared the gene expression values of the genes related to these two terms (Supplementary file 6). We are not surprised to observe that the genes of these two major contributing events of COVID-19 lung pathobiology are significantly overexpressed in lung samples compared to the rest of the SARS-CoV-2 infected cell types (Figure 6C-F, Supplementary Figure 4B-C). Particularly, the analyzed nasopharyngeal samples have very low expression values for the cytokine and inflammatory signaling genes (Figure 6C-F). Therefore, these observations are fueling the preexisting supposition of the roles of enhanced cytokine, and inflammatory signaling for worsening the disease condition in patients with SARS-CoV-2 infected lungs.

A differential gene expression profile was detected for the SARS-CoV-2 entry receptors/associated proteins in different infection models

Expression of receptor protein ACE2 and entry associated proteins such as- TMPRSS2, BSG, CTSL, DPP4 on the cell surface of the host is essential for the invasion of SARS-CoV-2 [62]. Moreover, ACE2 overexpression is thought to increase the infection potentiality of SARS-CoV-2 [68]. Furthermore, Kuba et al. demonstrated the potential role of ACE2 in SARS-CoV induced lung injury [69]. So, we ventured to check the gene expression levels of ACE2 and the other entry associated proteins in the different SARS-CoV-2 infected cells. Surprisingly, we observed that the *ACE2* gene is not expressed in high levels in lung samples as speculated (Figure 6G). However, gene expression levels of the other entry associated proteins were higher in lung samples (Figure 6G). Nonetheless, in few of the lung samples, the *TMPRSS2* gene was not expressed in higher amounts (Figure 6G). Interestingly, we have not detected any expression of *DPP4* gene within the Bangladeshi nasopharyngeal samples (Figure 6G).

Inflammatory immune responses were several folds higher in lungs than the nasopharynx of COVID-19 patients

From our previous observations, it was evident that COVID-19 patient's lung responds to the viral infection differently compared to the epithelial cells of nasopharynx. We then sought to figure out the specific genes and biological functions/signaling pathways which have this differential pattern. We achieved this by designing a multifactorial differential gene expression analysis using a generalized linear model (GLM) [44]; in which we compared the fold changes of every differentially expressed gene in nasopharyngeal and lung (GSE150316)

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. 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.

samples, to discover how many folds lung is alternatively expressing the genes than nasopharynx in COVID-19.

Firstly, we analyzed the suitability of the data for this design and observed no irregularities between the data used (Figure 7A-D). Moreover, upon this multifactorial differential gene expression analysis, we observed an acceptable common biological coefficient of variation; this variation decreases significantly as the expression values increases (Figure 7E). From the MA plot, we observed a very high amount of the significantly (p-value < 0.05) several fold upregulated and downregulated genes in lungs compared to nasopharyngeal samples (Figure 7F). We detected 807 upregulated and 298 downregulated genes in lungs compared to the nasopharyngeal samples (Supplementary file 7). Interestingly, we noticed the highly upregulated integrin and integrin signaling genes in lungs compared to the nasal samples (Figure 7G) which are consistent with our previous observations. Modulatory roles of integrins are well established in acute lung damages [70]. Similarly, aberrant expression of genes involved in integrin signaling can also provoke acute lung injuries, namely- *ADAM15* [71], *SDC1* [72], *CD14* [73], *CD47* [74], *CD9* [75], *HMGB1* [76], *ITA6* [77], and *ITAV* [78] etc. Therefore, SARS-CoV-2 infection induced deregulation of these genes might be contributing towards the worsening of the normal pathobiology and functionality of lungs in COVID-19.

We then performed functional enrichment analysis to hunt down the signaling pathways which are differentially expressed in lungs compared to the nasopharyngeal cells. These enrichment analyses revealed that biological functions such as viral process, antigen processing, and presentation etc. are highly upregulated, function such as- regulation of gene silencing by miRNA is found downregulated in lungs compared to the nasopharyngeal cells (Figure 8A). Furthermore, pathways that provide antiviral immunity such as apoptosis, phagosome, antigen processing and presentation, adaptive immune system, innate immune system, interferon signaling, different interleukin signaling, cytokine signaling in immune system etc. were highly upregulated in lungs compared to the nasopharyngeal samples (Figure 8B-D). Despite having the antiviral protective roles, hyperactivity from these pathways can significantly worsen the COVID-19 patient's overall lung functionality which can be further complicated with progressive and permanent lung damage.

Previously, it was reported that transcription factors can contribute to many inflammatory lung diseases [79, 80] which have similar lung characteristics observed in COVID-19. In this

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. 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.

context, we identified the highly expressed transcription factors in lungs by comparing their respective expression values in nasopharyngeal samples (Figure 8E). Among these, transcription factors such as- CBP [79], CEBP [81], NFAT [79], ATF3 [82], GATA6 [83], HDAC2 [84], TCF12 [85] etc. have significant roles in lung's overall functionality, acute lung injury and antiviral response mechanism in lungs.

SARS-CoV-2 integrates its proteins in regulating the host antiviral immune responses

As we have observed the differential host responses in COVID-19 nasopharyngeal samples, then we sought to interconnect the virus-host interplay in those host responses. We first analyzed how many of the virus interacting host proteins' genes reported by Gordon et al. [21] are differentially expressed in our reported nasopharyngeal samples. Only 51 genes of those proteins are found deregulated in our nasopharyngeal samples (Figure 9A). We then constructed a network interlinking the virus-host protein-protein interaction data from Gordon et al. [21] along with the deregulated genes from the nasopharyngeal samples (Figure 9B). Strikingly, we observed that most of the immune signaling related downregulated genes are directly or indirectly connected to the viral proteins (Figure 9B); this suggests the probable roles of the virus in the differential host responses in the COVID-19 affected patients.

Furthermore, we sought to establish the links between the viral proteins with integrin signaling associated genes by constructing a functional network with the viral-host protein-protein interaction data with the highly upregulated genes observed in lungs (from the comparison analysis between the lung and nasopharyngeal samples) (Figure 9C). From this constructed network, we observed that viral proteins such as ORF10, N, ORF9b, NSP7, NSP15, NSP5, M, NSP13, NSP2, NSP9, ORF8, ORF9c, NSP12, and NSP1 can directly or indirectly interact with the differentially expressed genes in lungs (Figure 9C), suggesting the putative mechanism behind the deregulated integrin signaling to promote the viral invasion in lungs.

Discussion

For a better understanding of the host-virus interaction in the SARS-CoV-2 pathogenesis, transcriptional responses of hosts play an enormous role. Moreover, population-specific disease induced host transcriptome might also correlate the susceptibility from COVID-19

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. 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.

with an individual's ethnicity [86]. In this context, we aimed to discover the representative host transcriptome response upon SARS-CoV-2 infection by performing and analyzing total RNA-seq from the nasopharyngeal samples of four COVID-19 positive Bangladeshi individuals. Also, we analyzed the genomic variations of the SARS-CoV-2 isolates from these patients to gain more insights into their probable origin. Moreover, we compared the transcriptome from different SARS-CoV-2 infection models, particularly, we compared the differential gene expression of the lung biopsy samples with the nasopharyngeal samples of ours to illustrate the possible molecular mechanisms behind the lung damages in severe COVID-19 patients.

Analyzing the variations within the sequenced isolates, we observed a very low frequency of missense amino acid variations within the proteins of SARS-CoV-2, particularly in the spike and ORF3a gene regions. Many concepts are there correlating the probable roles of variations with the COVID-19 disease severity [87, 88]. We did not observe any such variations within the spike region of our reported isolates; however, we recorded an unusual amount of 3'-UTR and 5'-UTR variations within these four isolates. Surprisingly, we spotted a very rare A>G mutation at position 10329 of ORF1ab gene in three of our four isolates; we also got another rare A>T variation at position 25505 of ORF3a gene in one of our isolates. Both of these variations are extremely rare amongst the other worldwide isolates (Table 1).

Phylogenetic analyses from Nextstrain portal suggested that our reported isolates are closely related to the Middle East Asian SARS-CoV-2 isolates; which is also very evident from our constructed neighbor-joining phylogenetic tree of 145 Bangladeshi isolates as our reported isolates are distinctly placed within the tree. This supports the idea of probable source of origin of these isolates from Middle-East Asia, while the rest of the Bangladeshi isolates are more related to the European isolates. As our isolates were obtained from the Chittagong area of Bangladesh, from where a large number of people recently immigrated to Middle-East (particularly Saudi Arabia) for work [89]; those immigrant people returning from the Middle-East during this pandemic might have brought these isolates into Bangladesh.

Previously, host transcriptional responses reported by Blanco-melo et al. [24] and Butler et al. [25] suggested a potential increase in the host antiviral immune responses such as- interferon signaling, interferon stimulated gene signaling, chemokine signaling, cytokine signaling etc.; however, Blanco-melo et al [90] also reported the presence of low IFN-I and IFN-III in COVID-19 patient's lung cells. We observed similar host immune responses, interferon, and

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. 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.

cytokine signaling in Bangladeshi COVID-19 patients too. Moreover, we also observed a stimulated innate immune response in our patients which was also reported for other COVID-19 patients [91].

Astoundingly, important signaling pathways those elicit antiviral immune responses such as- apoptosis [20], phagosome formation [92], antigen processing and presentation [93], Natural killer cell mediated cytotoxicity [94], and Toll-like receptor signaling [95] etc. were found downregulated in Bangladeshi COVID-19 patients. Also, pathways such as- HIF-1 response [96], PI3K-Akt signaling [97], and IL-17 signaling [98] etc. are also found deregulated which could assist the COVID-19 patients suffering from hypoxia, lung injury, and inflammation of the respiratory tract.

While we were comparing the nasopharyngeal cell's transcriptional responses with other SARS-CoV-2 infection models, we observed that lung cells elicited the immense cytokine and inflammatory responses against the invading viral pathogen. These overstimulated responses sometimes can do irreversible damages to the lungs [99]. This might shed insights into the COVID-19 disease severity when the viral infection progresses into the lungs.

Though an increased amount of ACE2 will facilitate the invasion of SARS-CoV-2, nonetheless, we observed a significant downregulation of ACE2 in lung cells. This phenomenon could backup the concept of ACE2 downregulation by SARS-CoV-2 itself after using it [100], thus reducing the organ protective roles of ACE2 [101] and resulting in progressive lung damages.

Integrins were reported important for the entry of SARS-CoV into the host cells [65], so it was speculated similar phenomenon might also be present in SARS-CoV-2. This idea is further intensified after the study by Sigrist et al. [63], who suggested the presence of an integrin-binding RGD motif in the spike of SARS-CoV-2. Surprisingly, upon the gene expression comparison between the different SARS-CoV-2 infected cells, we observed several folds upregulated expressions of genes related to the integrin signaling pathway in lung cells. This observation could support the idea of increased viral infections in lungs might be happening due to the overexpression of these probable attachment proteins. Also, the network analysis suggests a probable mechanism of upregulation of these proteins by the virus itself by the putative interactions through its proteins. Though more targeted studies should be undertaken for conclusive evidence supporting this phenomenon.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. 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.

Conclusion

In this study, we present the very first report of the host transcriptional response data from COVID-19 patients of the South-Asian region along with the SARS-CoV-2 isolates obtained from these patients. This data might provide newer insights into the host susceptibility from the perspective of ethnicity as well as will provide newer aspects of host responses against the virus in the different parts of the respiratory tract. However, a limited number of patient data is used here, but subsequent incorporation of more patient data from other parts of the world will significantly increase the understanding of this complex host-virus response in COVID-19, which will help in designing therapeutic interventions as well as in current clinical management of the patients.

Acknowledgement

We would like to thank all the authors who have kindly deposited and shared genome data on GISAID (https://www.gisaid.org/). We would also want to convey our gratitude to the developers maintaining the NextStrain novel coronavirus data portal (https://nextstrain.org/).

Competing Interests

The authors declare no competing interests.

Author Contributions

ABMMKI designed the workflow and conceived the project. AMAMZS performed sample collection and detection. RA, MSH, SMTK, and MSI performed the RNA sequencing and viral genome assembly. ABMMKI and MAAKK performed the RNA-seq data analysis, comparative genomics, and other bioinformatic analyses. MAAKK and ABMMKI wrote the manuscript. All authors read and approved the final manuscript. ABMMKI and MAAKK contributed equally to this work.

Funding

This project was not associated with any internal or external source of funding.

Data Availability

SARS-CoV-2 genomes are deposited at GISAID initiatives portal (https://www.gisaid.org/) with accessions- EPI_ISL_450340, EPI_ISL_450341, EPI_ISL_450342, EPI_ISL_450345.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. 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.

Raw RNA-seq data are deposited at NCBI-GEO (https://www.ncbi.nlm.nih.gov/geo/) under the accession- GSM4667504, GSM4667505, GSM4667506, GSM4667507. Additionally, publicly available data were utilized (Supplementary file 1). Analyses generated data are deposited as supplementary files.

Ethics statements

Patients' written-consents were taken before the collection of the samples.

References

1. Worldometer, *Coronavirus Cases*. 2020. p. 1-22.
2. Lu, R., et al., *Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding.* The Lancet, 2020. **395**(10224): p. 565-574.
3. NCBI-Gene, *Gene Links for Nucleotide (Select 1798174254) - Gene - NCBI*. 2020.
4. Jiang, S., L. Du, and Z. Shi, *An emerging coronavirus causing pneumonia outbreak in Wuhan, China: calling for developing therapeutic and prophylactic strategies*. Emerging Microbes & Infections, 2020. **9**(1): p. 275-277.
5. Liao, J., et al., *Epidemiological and clinical characteristics of COVID-19 in adolescents and young adults*. The Innovation, 2020. **1**(1): p. 100001.
6. Koh, J., et al., *Epidemiological and Clinical Characteristics of Cases During the Early Phase of COVID-19 Pandemic: A Systematic Review and Meta-Analysis*. Frontiers in Medicine, 2020. **7**(295).
7. Huang, C., et al., *Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China*. The Lancet, 2020. **395**(10223): p. 497-506.
8. Galiatsatos, P. *What Coronavirus Does to the Lungs*. 2020 [cited 2020 24 April]; Available from: https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/what-coronavirus-does-to-the-lungs.
9. Mao, L., et al., *Neurologic Manifestations of Hospitalized Patients With Coronavirus Disease 2019 in Wuhan, China*. JAMA Neurology, 2020.
10. Zheng, Y.-Y., et al., *COVID-19 and the cardiovascular system.* Nature Reviews Cardiology, 2020. **17**(5): p. 259-260.
11. Lauer, S.A., et al., *The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application.* Annals of Internal Medicine, 2020.
12. Lan, L., et al., *Positive RT-PCR Test Results in Patients Recovered From COVID-19.* JAMA, 2020. **323**(15): p. 1502-1503.
13. Chen, D., et al., *Recurrence of positive SARS-CoV-2 RNA in COVID-19: A case report.* International Journal of Infectious Diseases, 2020. **93**: p. 297-299.
14. Kirkcaldy, R.D., B.A. King, and J.T. Brooks, *COVID-19 and Postinfection Immunity: Limited Evidence, Many Remaining Questions.* JAMA, 2020. **323**(22): p. 2245-2246.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. 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.

15. Mehta, P., et al., *COVID-19: consider cytokine storm syndromes and immunosuppression*. The Lancet, 2020. **395**(10229): p. 1033-1034.
16. Yoshikawa, T., et al., *Dynamic innate immune responses of human bronchial epithelial cells to severe acute respiratory syndrome-associated coronavirus infection*. PloS one, 2010. **5**(1): p. e8729-e8729.
17. Ye, Q., B. Wang, and J. Mao, *The pathogenesis and treatment of the `Cytokine Storm' in COVID-19*. Journal of Infection, 2020.
18. Gu, J. and C. Korteweg, *Pathology and pathogenesis of severe acute respiratory syndrome*. The American journal of pathology, 2007. **170**(4): p. 1136-1147.
19. Schäfer, A. and R.S. Baric, *Epigenetic Landscape during Coronavirus Infection*. Pathogens (Basel, Switzerland), 2017. **6**(1): p. 8.
20. Fung, T.S. and D.X. Liu, *Human Coronavirus: Host-Pathogen Interaction*. Annual Review of Microbiology, 2019. **73**(1): p. 529-557.
21. Gordon, D.E., et al., *A SARS-CoV-2 protein interaction map reveals targets for drug repurposing*. Nature, 2020.
22. Zeberg, H. and S. Pääbo, *The major genetic risk factor for severe COVID-19 is inherited from Neandertals*. bioRxiv, 2020: p. 2020.07.03.186296.
23. Hassan, S.S., et al., *Molecular conservation and differential mutation on ORF3a gene in Indian SARS-CoV2 genomes*. Genomics, 2020. **112**(5): p. 3226-3237.
24. Blanco-Melo, D., et al., *Imbalanced Host Response to SARS-CoV-2 Drives Development of COVID-19*. Cell, 2020. **181**(5): p. 1036-1045.e9.
25. Butler, D.J., et al., *Shotgun Transcriptome and Isothermal Profiling of SARS-CoV-2 Infection Reveals Unique Host Responses, Viral Diversification, and Drug Interactions*. bioRxiv, 2020: p. 2020.04.20.048066.
26. Xiong, Y., et al., *Transcriptomic characteristics of bronchoalveolar lavage fluid and peripheral blood mononuclear cells in COVID-19 patients*. Emerging Microbes & Infections, 2020. **9**(1): p. 761-770.
27. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data*. Bioinformatics, 2014. **30**(15): p. 2114-2120.
28. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2*. Nature Methods, 2012. **9**(4): p. 357-359.
29. Li, D., et al., *MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph*. Bioinformatics, 2015. **31**(10): p. 1674-6.
30. Raney, B.J., et al., *Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser*. Bioinformatics, 2013. **30**(7): p. 1003-1005.
31. Ramírez, F., et al., *deepTools2: a next generation web server for deep-sequencing data analysis*. Nucleic Acids Research, 2016. **44**(W1): p. W160-W165.
32. Zhao, W.M., et al., *The 2019 novel coronavirus resource*. Yi Chuan, 2020. **42**(2): p. 212-221.
33. Corporation., M. *Microsoft Excel*. [cited 2020; Available from: https://office.microsoft.com/excel.
34. McLaren, W., et al., *The Ensembl Variant Effect Predictor*. Genome Biology, 2016. **17**(1): p. 122.
35. Barrett, T., et al., *NCBI GEO: archive for functional genomics data sets—update*. Nucleic Acids Research, 2012. **41**(D1): p. D991-D995.
36. Andrews, S., *FastQC: a quality control tool for high throughput sequence data*. 2010, Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. 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.

37. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq*. Bioinformatics, 2009. **25**(9): p. 1105-11.
38. Lander, E.S., et al., *Initial sequencing and analysis of the human genome*. Nature, 2001. **409**(6822): p. 860-921.
39. Hansen, K.D., S.E. Brenner, and S. Dudoit, *Biases in Illumina transcriptome sequencing caused by random hexamer priming*. Nucleic Acids Res, 2010. **38**(12): p. e131.
40. Hubbard, T.J.P., et al., *Ensembl 2007*. Nucleic acids research, 2007. **35**(suppl_1): p. D610-D617.
41. Liao, Y., G.K. Smyth, and W. Shi, *The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote*. Nucleic Acids Res, 2013. **41**(10): p. e108.
42. Anders, S. and W. Huber, *Differential expression analysis for sequence count data*. Genome Biology, 2010. **11**(10): p. R106-R106.
43. Kauffmann, A., R. Gentleman, and W. Huber, *arrayQualityMetrics—a bioconductor package for quality assessment of microarray data*. Bioinformatics, 2009. **25**(3): p. 415-416.
44. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data*. Bioinformatics (Oxford, England), 2010. **26**(1): p. 139-140.
45. Shu, Y. and J. McCauley, *GISAID: Global initiative on sharing all influenza data–from vision to reality*. Eurosurveillance, 2017. **22**(13): p. 30494.
46. Katoh, K. and D.M. Standley, *MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability*. Molecular Biology and Evolution, 2013. **30**(4): p. 772-780.
47. Kumar, S., G. Stecher, and K. Tamura, *MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets*. Mol Biol Evol, 2016. **33**(7): p. 1870-4.
48. Perez-Llamas, C. and N. Lopez-Bigas, *Gitools: analysis and visualisation of genomic data using interactive heat-maps*. PloS one, 2011. **6**(5).
49. Ashburner, M., et al., *Gene ontology: tool for the unification of biology*. Nature genetics, 2000. **25**(1): p. 25-29.
50. Huang, R., et al., *The NCATS BioPlanet – An Integrated Platform for Exploring the Universe of Cellular Signaling Pathways for Toxicology, Systems Biology, and Chemical Genomics*. Frontiers in Pharmacology, 2019. **10**(445).
51. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes*. Nucleic acids research, 2000. **28**(1): p. 27-30.
52. Jassal, B., et al., *The reactome pathway knowledgebase*. Nucleic Acids Res, 2020. **48**(D1): p. D498-d503.
53. Benjamini, Y. and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing*. Journal of the Royal statistical society: series B (Methodological), 1995. **57**(1): p. 289-300.
54. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks*. Genome Res, 2003. **13**(11): p. 2498-504.
55. Szklarczyk, D., et al., *STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets*. Nucleic acids research, 2019. **47**(D1): p. D607-D613.
56. Hu, T., et al., *Two adjacent mutations on the dimer interface of SARS coronavirus 3C-like protease cause different conformational changes in crystal structure*. Virology, 2009. **388**(2): p. 324-334.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. 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.

57. Muramatsu, T., et al., *SARS-CoV 3CL protease cleaves its C-terminal autoprocessing site by novel subsite cooperativity.* Proceedings of the National Academy of Sciences, 2016. **113**(46): p. 12997-13002.
58. Huang, C., et al., *3C-like Proteinase from SARS Coronavirus Catalyzes Substrate Hydrolysis by a General Base Mechanism*. Biochemistry, 2004. **43**: p. 4568-74.
59. Hadfield, J., et al., *Nextstrain: real-time tracking of pathogen evolution.* Bioinformatics, 2018. **34**(23): p. 4121-4123.
60. Zhang, X., et al., *Viral and host factors related to the clinical outcome of COVID-19.* Nature, 2020. **583**(7816): p. 437-440.
61. Townsend, M.J., T.K. Kyle, and F.C. Stanford, *Outcomes of COVID-19: disparities in obesity and by ethnicity/race*. International Journal of Obesity, 2020.
62. Sungnak, W., et al., *SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes*. Nature Medicine, 2020.
63. Sigrist, C.J., A. Bridge, and P. Le Mercier, *A potential role for integrins in host cell entry by SARS-CoV-2.* Antiviral research, 2020. **177**: p. 104759-104759.
64. Tresoldi, I., et al., *SARS-COV-2 and infectivity: Possible increase in infectivity associated to integrin motif expression.* Journal of Medical Virology. **n/a**(n/a).
65. Hänel, K., et al., *Solution structure of the X4 protein coded by the SARS related coronavirus reveals an immunoglobulin like fold and suggests a binding activity to integrin I domains.* Journal of Biomedical Science, 2006. **13**(3): p. 281-293.
66. Colafrancesco, S., et al., *Targeting the Immune System for Pulmonary Inflammation and Cardiovascular Complications in COVID-19 Patients.* Frontiers in Immunology, 2020. **11**(1439).
67. Pedersen, S.F. and Y.C. Ho, *SARS-CoV-2: a storm is raging.* J Clin Invest, 2020. **130**(5): p. 2202-2205.
68. Xu, H., et al., *High expression of ACE2 receptor of 2019-nCoV on the epithelial cells of oral mucosa.* International Journal of Oral Science, 2020. **12**(1): p. 8.
69. Kuba, K., et al., *A crucial role of angiotensin converting enzyme 2 (ACE2) in SARS coronavirus–induced lung injury.* Nature Medicine, 2005. **11**(8): p. 875-879.
70. Sheppard, D., *Modulation of Acute Lung Injury by Integrins.* Proceedings of the American Thoracic Society, 2012. **9**(3): p. 126-129.
71. Sun, C., et al., *ADAM15 deficiency attenuates pulmonary hyperpermeability and acute lung injury in lipopolysaccharide-treated mice.* American journal of physiology. Lung cellular and molecular physiology, 2013. **304**(3): p. L135-L142.
72. Parimon, T., et al., *Syndecan-1 promotes lung fibrosis by regulating epithelial reprogramming through extracellular vesicles.* JCI insight, 2019. **5**(17): p. e129359.
73. Anas, A., T. van der Poll, and A.F. de Vos, *Role of CD14 in lung inflammation and infection.* Critical care (London, England), 2010. **14**(2): p. 209-209.
74. Su, X., et al., *CD47 deficiency protects mice from lipopolysaccharide-induced acute lung injury and Escherichia coli pneumonia.* Journal of immunology (Baltimore, Md. : 1950), 2008. **180**(10): p. 6947-6953.
75. Takeda, Y., et al., *Preventive Role of Tetraspanin CD9 in Systemic Inflammation of Chronic Obstructive Pulmonary Disease.* American Journal of Respiratory Cell and Molecular Biology, 2015. **53**(6): p. 751-760.
76. Ding, J., X. Cui, and Q. Liu, *Emerging role of HMGB1 in lung diseases: friend or foe.* Journal of cellular and molecular medicine, 2017. **21**(6): p. 1046-1057.
77. Chen, H., et al., *Mechanosensing by the α6-integrin confers an invasive fibroblast phenotype and mediates lung fibrosis.* Nature communications, 2016. **7**: p. 12564-12564.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. 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.

78. Luzina, I.G., et al., *Regulation of pulmonary inflammation and fibrosis through expression of integrins αVβ3 and αVβ5 on pulmonary T lymphocytes*. Arthritis & Rheumatism, 2009. **60**(5): p. 1530-1539.
79. Rahman, I. and W. MacNee, *Role of transcription factors in inflammatory lung diseases*. Thorax, 1998. **53**(7): p. 601-612.
80. Fan, J., R.D. Ye, and A.B. Malik, *Transcriptional mechanisms of acute lung injury*. Am J Physiol Lung Cell Mol Physiol, 2001. **281**(5): p. L1037-50.
81. Cassel, T.N. and M. Nord, *C/EBP transcription factors in the lung epithelium*. American Journal of Physiology-Lung Cellular and Molecular Physiology, 2003. **285**(4): p. L773-L781.
82. Qian, L., et al., *Activating transcription factor 3 (ATF3) protects against lipopolysaccharide-induced acute lung injury via inhibiting the expression of TL1A.* Journal of Cellular Physiology, 2017. **232**(12): p. 3727-3734.
83. Bruno, M.D., et al., *GATA-6 activates transcription of surfactant protein A.* J Biol Chem, 2000. **275**(2): p. 1043-9.
84. Barnes, P.J., *Role of HDAC2 in the pathophysiology of COPD.* Annu Rev Physiol, 2009. **71**: p. 451-64.
85. Li, L., et al., *Angiopoietin-like 4 Increases Pulmonary Tissue Leakiness and Damage during Influenza Pneumonia.* Cell Rep, 2015. **10**(5): p. 654-663.
86. Shi, Y., et al., *COVID-19 infection: the perspectives on immune responses.* Cell Death & Differentiation, 2020. **27**(5): p. 1451-1454.
87. Biswas, S.K. and S.R. Mudi, *Genetic variation in SARS-CoV-2 may explain variable severity of COVID-19.* Medical hypotheses, 2020. **143**: p. 109877-109877.
88. Korber, B., et al., *Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus*. Cell, 2020.
89. Express, T.F. *Manpower export from Chittagong region rises in 2017*. 2018; Available from: https://thefinancialexpress.com.bd/economy/bangladesh/manpower-export-from-276.
90. Blanco-Melo, D., et al., *Imbalanced host response to SARS-CoV-2 drives development of COVID-19.* Cell, 2020.
91. Liao, M., et al., *Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19.* Nature Medicine, 2020. **26**(6): p. 842-844.
92. Ahmad, L., S. Mostowy, and V. Sancho-Shimizu, *Autophagy-Virus Interplay: From Cell Biology to Human Disease.* Frontiers in cell and developmental biology, 2018. **6**: p. 155-155.
93. Yewdell, J.W. and A.B. Hill, *Viral interference with antigen presentation.* Nat Immunol, 2002. **3**(11): p. 1019-25.
94. Brandstadter, J.D. and Y. Yang, *Natural killer cell responses to viral infection.* Journal of innate immunity, 2011. **3**(3): p. 274-279.
95. Olejnik, J., A.J. Hume, and E. Mühlberger, *Toll-like receptor 4 in acute viral infection: Too much of a good thing.* PLoS pathogens, 2018. **14**(12): p. e1007390-e1007390.
96. Chen, S. and N. Sang, *Hypoxia-Inducible Factor-1: A Critical Player in the Survival Strategy of Stressed Cells.* Journal of cellular biochemistry, 2016. **117**(2): p. 267-278.
97. Alexiou, K., et al., *Relaxin protects rat lungs from ischemia-reperfusion injury via inducible NO synthase: role of ERK-1/2, PI3K, and forkhead transcription factor FKHRL1*. PloS one, 2013. **8**(9): p. e75592-e75592.
98. Gurczynski, S.J. and B.B. Moore, *IL-17 in the lung: the good, the bad, and the ugly.* American journal of physiology. Lung cellular and molecular physiology, 2018. **314**(1): p. L6-L16.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. 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.

99. Atamas, S.P., S.P. Chapoval, and A.D. Keegan, *Cytokines in chronic respiratory diseases*. F1000 biology reports, 2013. **5**: p. 3-3.
100. Ciulla, M.M., *SARS-CoV-2 downregulation of ACE2 and pleiotropic effects of ACEIs/ARBs*. Hypertension Research, 2020.
101. Cheng, H., Y. Wang, and G.-Q. Wang, *Organ-protective effect of angiotensin-converting enzyme 2 and its effect on the prognosis of COVID-19.* Journal of Medical Virology, 2020. **92**(7): p. 726-730.

List of Tables

Table 1: Observed variations within the four Bangladeshi SARS-CoV-2 isolates reported in this study.

Genomic position: Variation	Variation Type	Associated genomic region	Protein : amino acid change	Frequency in the four isolates used in this study	Frequency in the other Bangladeshi isolates	Frequency in the isolates from rest of the world
1: ATTAAAGGTTTA>-	Intergenic variant	5'UTR	-	1	-	-
1: ATTAAAGGTTTATA>-	Intergenic variant	5'UTR	-	1	-	-
2: T>TA	Intergenic variant	5'UTR	-	1	0	0
2: T>TTTCAAAGATCAAGTCA	Intergenic variant	5'UTR	-	1	0	0
4: A>T	Intergenic variant	5'UTR	-	1	0	58
7: G>C	Intergenic variant	5'UTR	-	1	0	16
9: T>TTTTCGC	Intergenic	5'UTR	-	1	0	0

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. 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.

Genomic position: Variation	Variation Type	Associated genomic region	Protein : amino acid change	Frequency in the four isolates used in this study	Frequency in the other Bangladeshi isolates	Frequency in the isolates from rest of the world
	variant					
12: A>T	Intergenic variant	5'UTR	-	1	0	22
13: T>C	Intergenic variant	5'UTR	-	1	0	36
280: C>T	Synonymous variant	orf1ab	5V	1	0	5
601: C>T	Synonymous variant	orf1ab	112G	1	1	6
8782: C>T	Synonymous variant	orf1ab	2839S	4	1	3012
10323: A>G	Missense variant	orf1ab	3353K>R	1	5	154
10329: A>G	Missense variant	orf1ab	3355D>G	3	1	1
10870: G>T	Synonymous variant	orf1ab	3535L	1	0	27
12119: C>T	Missense variant	orf1ab	3952P>S	1	0	8
15324: C>T	Synonymous variant	orf1ab	5020N	3	5	818
19414: G>A	Missense variant	orf1ab	6384V>I	1	0	0
22468: G>T	Synonymous variant	S	302T	4	1	99
23320: C>T	Synonymous variant	S	586D	1	0	2
25505: A>T	Missense variant	ORF3a	38Q>L	1	0	2
28087: C>T	Missense variant	ORF8	65A>V	1	0	23
28144: T>C	Missense	ORF8	84L>S	4	1	3050

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. 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.

Genomic position: Variation	Variation Type	Associated genomic region	Protein : amino acid change	Frequency in the four isolates used in this study	Frequency in the other Bangladeshi isolates	Frequency in the isolates from rest of the world
	variant					
28878: G>A	Missense variant	N	202S>N	4	1	253
29392: G>T	Missense variant	N	373K>N	2	2	6
29742: G>A	Downstream gene variant	3'UTR	-	4	1	21
29856: T>A	Intergenic variant	3'UTR	-	1	0	6
29858: T>A	Intergenic variant	3'UTR	-	1	0	5
29864: GAATGACAAAAAAAAAAAAAAAAAAAAA>G	Intergenic variant	3'UTR	-	1	0	0
29864: GAATGACAAAAAAAAAAAAAAAAAAAAA>T	Intergenic variant	3'UTR	-	1	0	0
29870: CAAAAAAAAAAAAAAAAAAAAAAAAAA>C	Intergenic variant	3'UTR	-	1	1	-
29870: C>G	Intergenic variant	3'UTR	-	1	0	3
29872: A>T	Intergenic variant	3'UTR	-	1	0	12
29873: A>C	Intergenic variant	3'UTR	-	1	0	3
29874: A>G	Intergenic variant	3'UTR	-	1	0	12
29875: A>G	Intergenic variant	3'UTR	-	1	1	5
29878: A>T	Intergenic	3'UTR	-	1	0	3

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. 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.

Genomic position: Variation	Variation Type	Associated genomic region	Protein : amino acid change	Frequency in the four isolates used in this study	Frequency in the other Bangladeshi isolates	Frequency in the isolates from rest of the world
	variant					
29880: A>G	Intergenic variant	3'UTR	-	1	2	5
29882: A>G	Intergenic variant	3'UTR	-	1	0	13
29883: A>T	Intergenic variant	3'UTR	-	1	0	8
29884: A>C	Intergenic variant	3'UTR	-	1	0	10
29885: A>G	Intergenic variant	3'UTR	-	1	1	11
29886: A>T	Intergenic variant	3'UTR	-	1	1	5
29887: A>G	Intergenic variant	3'UTR	-	1	0	11
29888: A>T	Intergenic variant	3'UTR	-	1	1	4
29890: A>G	Intergenic variant	3'UTR	-	1	1	6
29891: A>G	Intergenic variant	3'UTR	-	1	0	15
29892: A>G	Intergenic variant	3'UTR	-	1	0	13
29893: A>G	Intergenic variant	3'UTR	-	1	0	15
29895: A>T	Intergenic variant	3'UTR	-	2	0	6
29896: A>G	Intergenic variant	3'UTR	-	1	0	7
29896: A>C	Intergenic variant	3'UTR	-	1	0	3
29897: A>G	Intergenic	3'UTR	-	2	0	4

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. 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.

Genomic position: Variation	Variation Type	Associated genomic region	Protein : amino acid change	Frequency in the four isolates used in this study	Frequency in the other Bangladeshi isolates	Frequency in the isolates from rest of the world
	variant					
29898: A>G	Intergenic variant	3'UTR	-	2	0	6
29898: A>T	Intergenic variant	3'UTR	-	1	0	5
29900: A>G	Intergenic variant	3'UTR	-	1	0	10
29901: AAA>A	Intergenic variant	3'UTR	-	1	0	-
29901: A>G	Intergenic variant	3'UTR	-	2	0	5
29903: A>GCCGTCGT	Intergenic variant	3'UTR	-	1	0	-
29903: A>GCGTCGTGT	Intergenic variant	3'UTR	-	1	0	-

Figure Legends

Figure 1: Genomic information of the sequenced SARS-CoV-2 isolates. **A.** Genome coverage normalized density map for the four sequenced SARS-CoV-2 isolates. **B.** Pie-chart illustrating the different types of variations found within these four isolates. **C.** Genome location-wise representation of the mutations and their associated frequency. **D.** Isolate-wise variation information. **E.** Gene-wise amount and type of mutations. **F.** Annotated impacts of the different mutations (only those are shown which have frequencies more than 1). **G.** Frequencies of selected unique mutations observed in these isolates.

Figure 2: Phylogenetic tree of Bangladeshi SARS-CoV-2 isolates. Neighbor-joining tree using MEGA tools. Isolates reported in this study are indicated with a red arrow. The evolutionary history was inferred using the Neighbor-Joining method. The optimal tree with the sum of branch length = 0.01403419 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. 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 branches. The evolutionary distances were computed using the Maximum Composite Likelihood method and are in the units of the number of base substitutions per site. This analysis involved 145 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. All positions with less than 95% site coverage were eliminated, i.e., fewer than 5% alignment gaps, missing data, and ambiguous bases were allowed at any position (partial deletion option). There was a total of 29827 positions in the final dataset. Values represent bootstrap numbers (%).

Figure 3: A. Variance plot. This plots the standard deviation of the transformed data, across samples, against the mean, using the variance stabilizing transformation. The vertical axis in the plots is the square root of the variance over all samples. **B.** Sample to sample distance plot. A heatmap of distance matrix providing an overview of similarities and dissimilarities between samples. Clustering is based on the distances between the rows/columns of the distance matrix. **C.** Principal component analysis plot. Samples are in the 2D plane spanned by their first two principal components. **D.** Clustered heatmap of the $\log_2$ converted normalized count matrix RNA-seq reads, top 50 genes, of nasopharyngeal samples. **E.** Normalized Log_2 read counts of the genes encoding SARS-CoV-2 receptor and entry associated proteins. Enrichment analysis and comparison between deregulated genes and the genes of some selected processes in SARS-CoV-2 infected nasopharyngeal samples and SARS-CoV-2 infected lung biopsy samples using **F.** GOBP module, **G.** KEGG pathway, **H.** Bioplanet pathway module. Selected significant terms are represented in heatmaps. Significance of enrichment in terms of the adjusted p-value (< 0.05) is represented in color-coded P-value scale for all heatmaps; Color towards red indicates higher significance and color towards yellow indicates less significance, while grey means non-significant. Normalized Log_2 converted read counts are considered as the expression values of the genes and represented in a color-coded scale; Color towards red indicating higher expression, while color towards green indicating little to no expression.

Figure 4: Comparison of the gene expression patterns in different SARS-CoV-2 infection models. **A.** Venn-diagram showing the observed deregulated genes (with their respective control) in the different cell types. Enrichment analysis and comparison between deregulated genes in different SARS-CoV-2 infection models using **B.** GOBP module, **C.** Bioplanet pathway module, **D.** KEGG pathway module. Selected significant terms are represented in heatmaps. Color scale/schemes are similar to Figure 3.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. 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 5: A. Variance plot, **B.** Sample to sample distance plot, **C.** Principal component analysis plot, **D.** Clustered heatmap of the count matrix of the normalized RNA-seq reads of different SARS-CoV-2 infection samples using to 50 genes. **E.** Gene expression heatmap showing global gene expression profiles in the individual infected samples of the various infection system. Heatmap is clustered based on Pearson's distance with genes that vary across the sample, leaving out genes that do not vary significantly.

Figure 6: Heatmaps representing the sample level absolute expression of Integrin related genes **A.** across the different SARS-CoV-2 infection models, **B.** in only nasopharyngeal samples and lung samples; Cytokine signaling related genes **C.** across the different SARS-CoV-2 infection models, **D.** in only nasopharyngeal samples and lung samples; and Inflammation related genes **E.** across the different SARS-CoV-2 infection models, **F.** in only nasopharyngeal samples and lung samples; **G.** Expression profiles of genes encoding SARS-CoV-2 receptor and entry associated proteins. Normalized (DESeq2) Log_2 converted read counts are considered as the expression values of the genes and represented in a color-coded scale; Color towards red indicating higher expression, while color towards green indicating little to no expression.

Figure 7: A. Variance plot, **B.** Sample to sample distance plot, **C.** Principal component analysis plot, **D.** Clustered heatmap of the count matrix of the normalized RNA-seq reads (top 50 genes) of the SARS-CoV-2 infected nasopharyngeal and lung samples. **E.** Common dispersion plot or the biological coefficient of variation plot. Here we are estimating the dispersion. The square root of the common dispersion gives the coefficient of variation of biological variation. Here the coefficient of biological variation is around 0.8. **F.** MA plot. Plot log-fold change against log-counts per million, with DE genes are highlighted. The blue lines indicate 2-fold changes. Red and blue points indicate genes with P-value less than 0.05. **G.** Expression profiles of genes encoding Integrins. Log_2 (fold change) values are considered as the expression values of the genes and represented in a color-coded scale; Color towards red indicating higher expression, while color towards green indicating little to no expression.

Figure 8: Enrichment analysis and comparison between deregulated genes and the genes of some selected processes in SARS-CoV-2 infected nasopharyngeal samples versus SARS-CoV-2 infected lung biopsy samples using **A.** GOBP module, **B.** KEGG pathway, **C.** Bioplanet pathway module, **D.** Reactome pathway module. Selected significant terms are represented in heatmaps. Color schemes are similar to Figure 3. For individual processes,

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. 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.

blue means presence (significantly differentially expressed gene) while grey means absence (not significantly differentially expressed genes for this module for this experimental condition).

Figure 9: A. Venn diagram showing the commonly deregulated genes between deregulated genes in our nasopharyngeal samples and Gordon et al. reported viral protein-interacting high confidence host proteins. Network representing the interactions between genes in **B.** Deregulated genes in Bangladeshi nasopharyngeal samples along with SARS-CoV-2 proteins and Gordon et al. described viral interacting host proteins, and **C.** Differentially expressed Integrin related genes in lungs compared to the nasal samples along with SARS-CoV-2 proteins and Gordon et al. described viral interacting host proteins. Hexagon, ellipse, rounded rectangle represents viral proteins, process-related genes, and proteins that interact with viral proteins, respectively. Expression values of the genes and represented in a color-coded scale. Color towards red indicating higher expression, while color towards green indicating little to no expression.

Supplementary Figure Legends

Supplementary Figure 1: Snapshot of Nextstrain data portal showing the phylogenetic relationship of two SARS-CoV-2 isolates used in this study. Isolates of this study are indicated using a red arrow.

Supplementary Figure 2: Deregulated genes of selected terms from Figure 3 in different SARS-CoV-2 infection systems. Genes of selected significant terms are represented here. For individual processes, blue means presence (differentially expressed gene of the module term) while grey means absence (not differentially expressed in the experimental condition in that module term). Processes in the green, blue, red color background represent KEGG, Bioplanet, GOBP enriched terms, respectively.

Supplementary Figure 3: Deregulated genes of selected terms from Figure 3 in different SARS-CoV-2 infection systems. For individual processes, blue means presence (differentially expressed gene of the module term) while grey means absence (not differentially expressed in the experimental condition in that module term). Processes in the green, blue, red color background represent KEGG, Bioplanet, GOBP enriched terms, respectively.

Supplementary Figure 4: Expanded view of the heatmaps A, B, C of Figure 6.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. 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.

List of supplementary files

Supplementary file 1: Sources of the data used in this study.

Supplementary file 2: Per base sequence quality reports of the generated RNA-seq reads of the four COVID-19 infected nasal samples used in this study.

Supplementary file 3: Isolate-wise variation information of the four SARS-CoV-2 isolates used in this study.

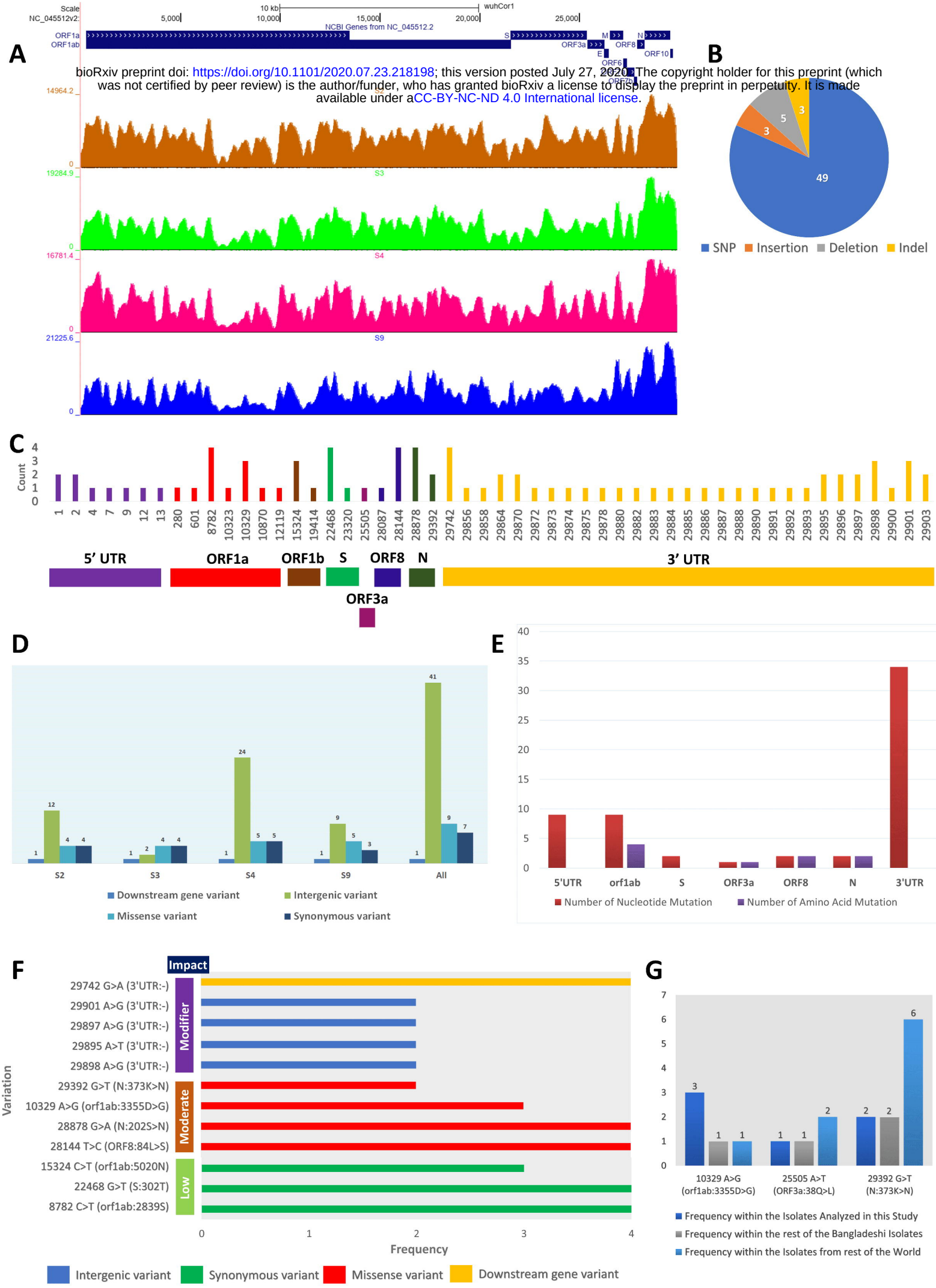
Supplementary file 4: Differentially expressed genes found in the four Nasal samples of Bangladeshi COVID-19 patient.

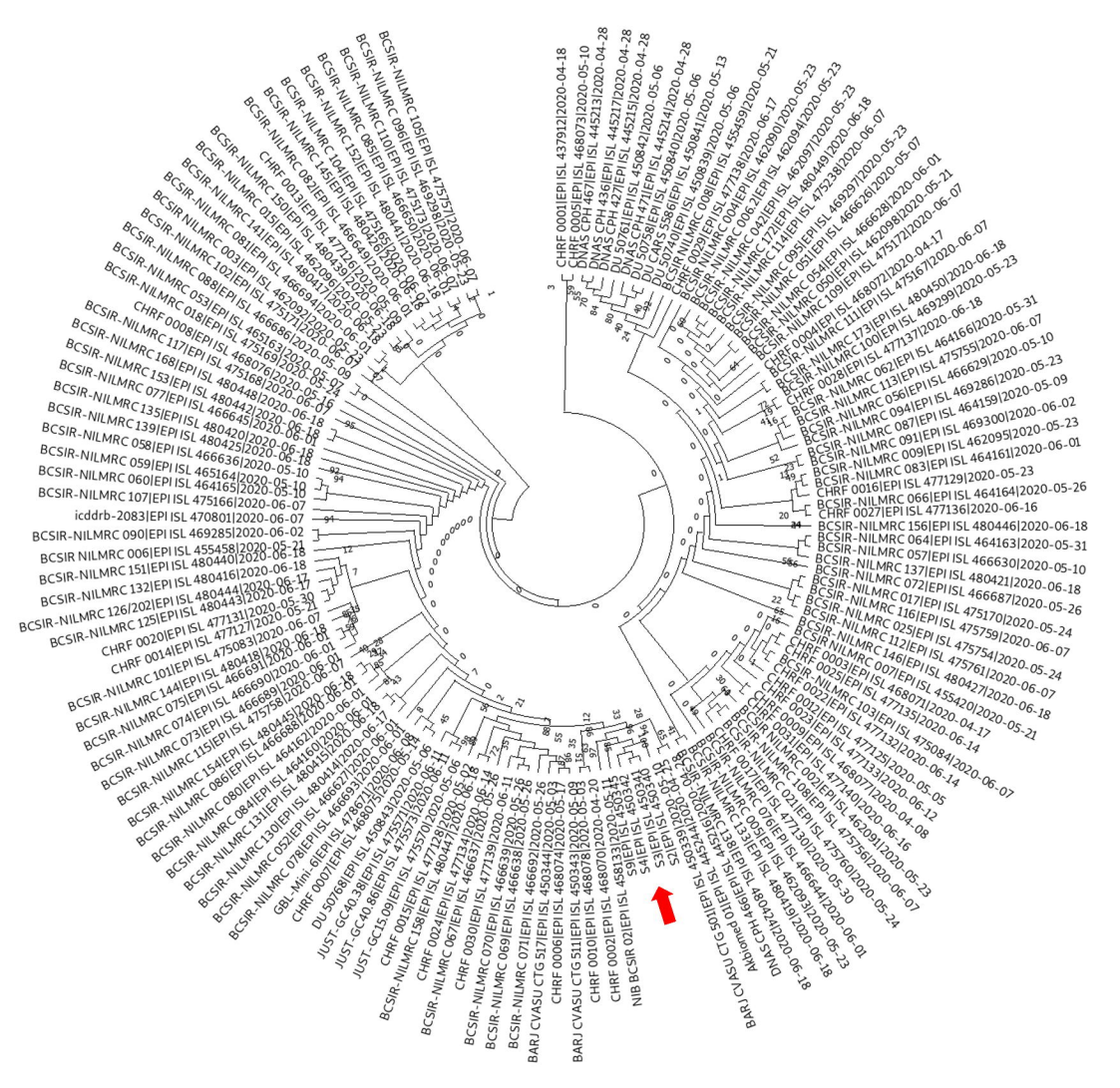
Supplementary file 5: Differentially expressed genes in different SARS-CoV-2 infected cell types.

Supplementary file 6: Genes and associated terms used for filtering the expression values used in Figure 6.

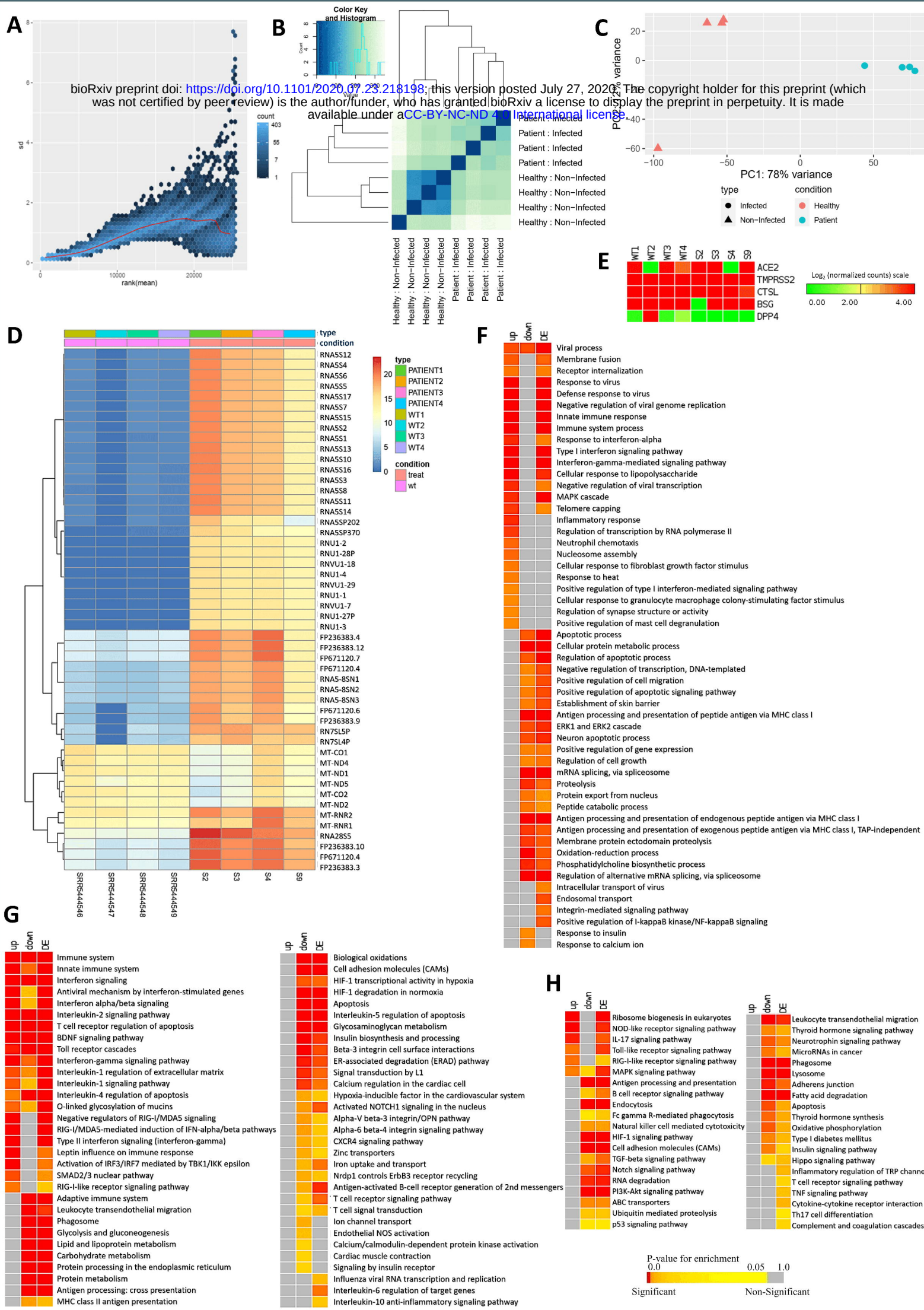
Supplementary file 7: Differentially expressed genes in SARS-CoV-2 infected Lungs compared to the Bangladeshi Nasal samples used in this study.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. 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.

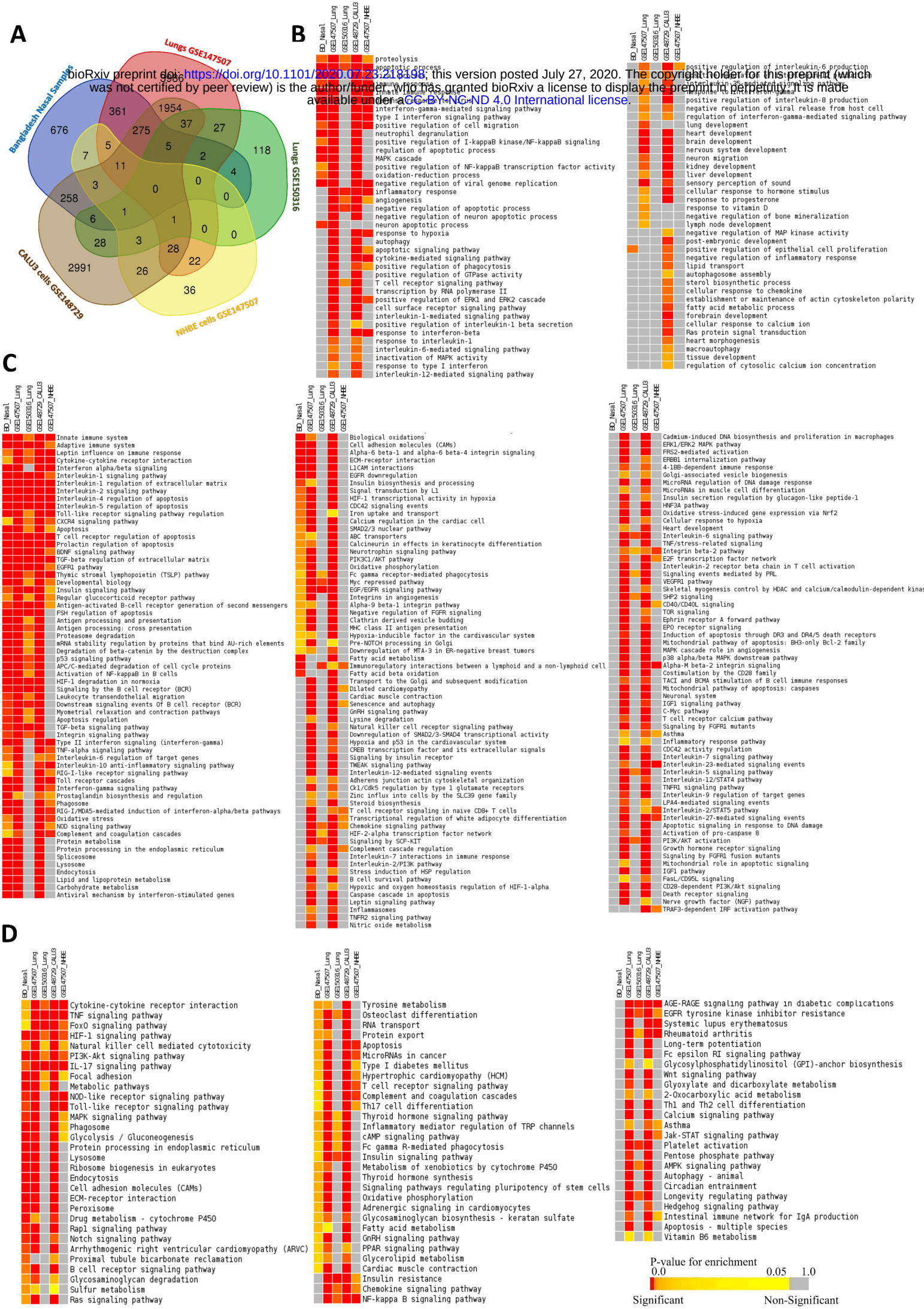



CHRF 0001|EPI ISL 437912|2020-04-18
DNAS CPH 0005|EPI ISL 468073|2020-05-10
DNAS CPH 467|EPI ISL 445213|2020-04-28
DNAS CPH 436|EPI ISL 445217|2020-04-28
DU 5076|EPI ISL 445215|2020-04-28
DU 5078|EPI ISL 450842|2020-05-06
DU CPH 471|EPI ISL 445214|2020-04-28
DU CARS 5586|EPI ISL 450840|2020-05-06
DU 5074|EPI ISL 450839|2020-05-13
BCSIR-NILMRC 008|EPI ISL 455459|2020-05-21
BCSIR-NILMRC 006|EPI ISL 462890|2020-05-23
BCSIR-NILMRC 042|EPI ISL 462094|2020-05-23
BCSIR-NILMRC 172|EPI ISL 462097|2020-05-23
BCSIR-NILMRC 085|EPI ISL 480449|2020-05-23
BCSIR-NILMRC 114|EPI ISL 475238|2020-06-07
BCSIR-NILMRC 054|EPI ISL 469297|2020-06-07
BCSIR-NILMRC 050|EPI ISL 466628|2020-06-07
BCSIR-NILMRC 109|EPI ISL 475167|2020-06-07
BCSIR-NILMRC 111|EPI ISL 469299|2020-05-23
BCSIR-NILMRC 173|EPI ISL 468071|2020-04-17
BCSIR-NILMRC 100|EPI ISL 480450|2020-06-18
CHRF 0004|EPI ISL 477137|2020-06-07
BCSIR-NILMRC 062|EPI ISL 464166|2020-05-31
CHRF 0028|EPI ISL 475755|2020-06-07
BCSIR-NILMRC 113|EPI ISL 466629|2020-05-10
BCSIR-NILMRC 056|EPI ISL 469286|2020-05-23
BCSIR-NILMRC 094|EPI ISL 464159|2020-05-09
BCSIR-NILMRC 087|EPI ISL 469300|2020-06-02
BCSIR-NILMRC 091|EPI ISL 462095|2020-05-23
BCSIR-NILMRC 009|EPI ISL 464161|2020-06-01
BCSIR-NILMRC 083|EPI ISL 477129|2020-05-23
CHRF 0016|EPI ISL 464164|2020-05-26
BCSIR-NILMRC 066|EPI ISL 477136|2020-06-16
CHRF 0027|EPI ISL 480446|2020-06-18
BCSIR-NILMRC 156|EPI ISL 464163|2020-05-31
BCSIR-NILMRC 064|EPI ISL 466630|2020-05-10
BCSIR-NILMRC 057|EPI ISL 480421|2020-06-18
BCSIR-NILMRC 137|EPI ISL 466687|2020-05-26
BCSIR-NILMRC 072|EPI ISL 475170|2020-05-24
BCSIR-NILMRC 017|EPI ISL 475759|2020-06-07
BCSIR-NILMRC 116|EPI ISL 475754|2020-05-24
BCSIR-NILMRC 025|EPI ISL 475761|2020-06-07
BCSIR-NILMRC 112|EPI ISL 480427|2020-06-18
BCSIR-NILMRC 146|EPI ISL 455420|2020-05-21
CHRF 0003|EPI ISL 468071|2020-04-17
BCSIR-NILMRC 071|EPI ISL 477135|2020-06-14
CHRF 0025|EPI ISL 475084|2020-06-07
CHRF 0022|EPI ISL 477132|2020-06-14
CHRF 0023|EPI ISL 477133|2020-06-12
CHRF 0012|EPI ISL 468071|2020-04-08
CHRF 0009|EPI ISL 475760|2020-05-05
CHRF 0031|EPI ISL 462091|2020-06-16
BCSIR-NILMRC 021|EPI ISL 475756|2020-05-23
BCSIR-NILMRC 095|EPI ISL 475130|2020-05-30
BCSIR-NILMRC 076|EPI ISL 466644|2020-06-01
BCSIR-NILMRC 133|EPI ISL 480419|2020-05-24
BCSIR-NILMRC 106|EPI ISL 462093|2020-05-23
BCSIR-NILMRC 017|EPI ISL 462098|2020-06-18
BCSIR-NILMRC 005|EPI ISL 462092|2020-05-23
BCSIR-NILMRC 121|EPI ISL 477138|2020-06-17
BCSIR-NILMRC 004|EPI ISL 455459|2020-05-21
BCSIR-NILMRC 042|EPI ISL 462094|2020-05-23
BCSIR-NILMRC 006|EPI ISL 462890

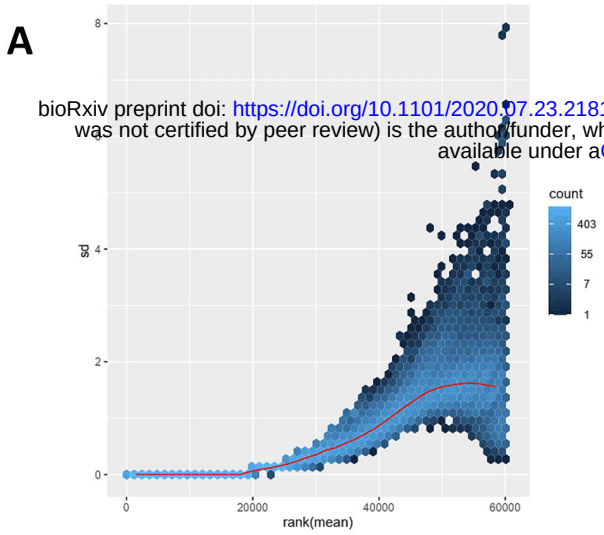
bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

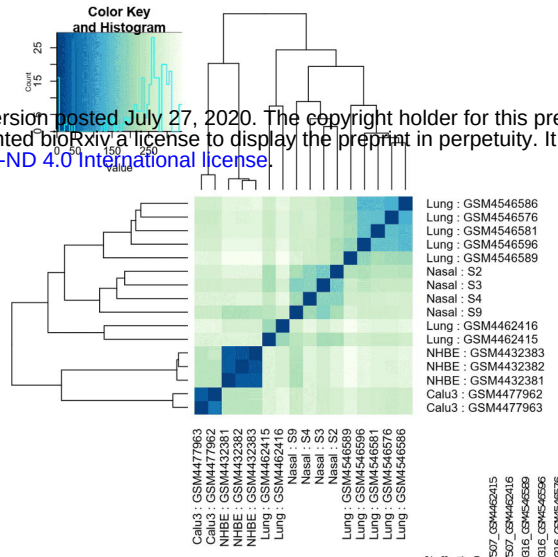


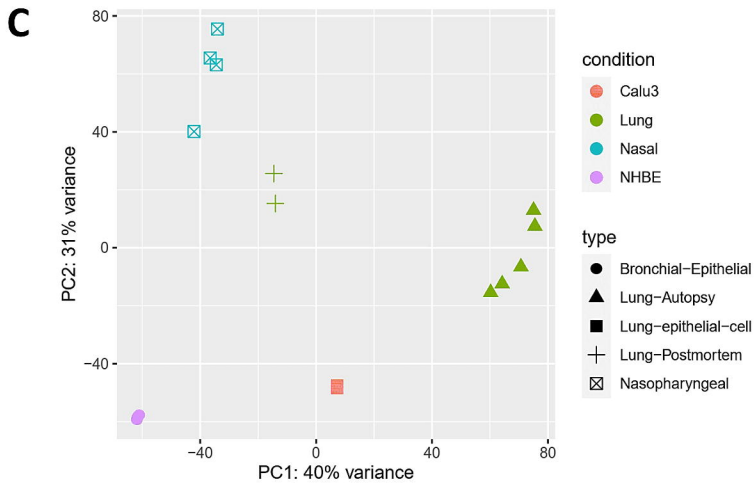
bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

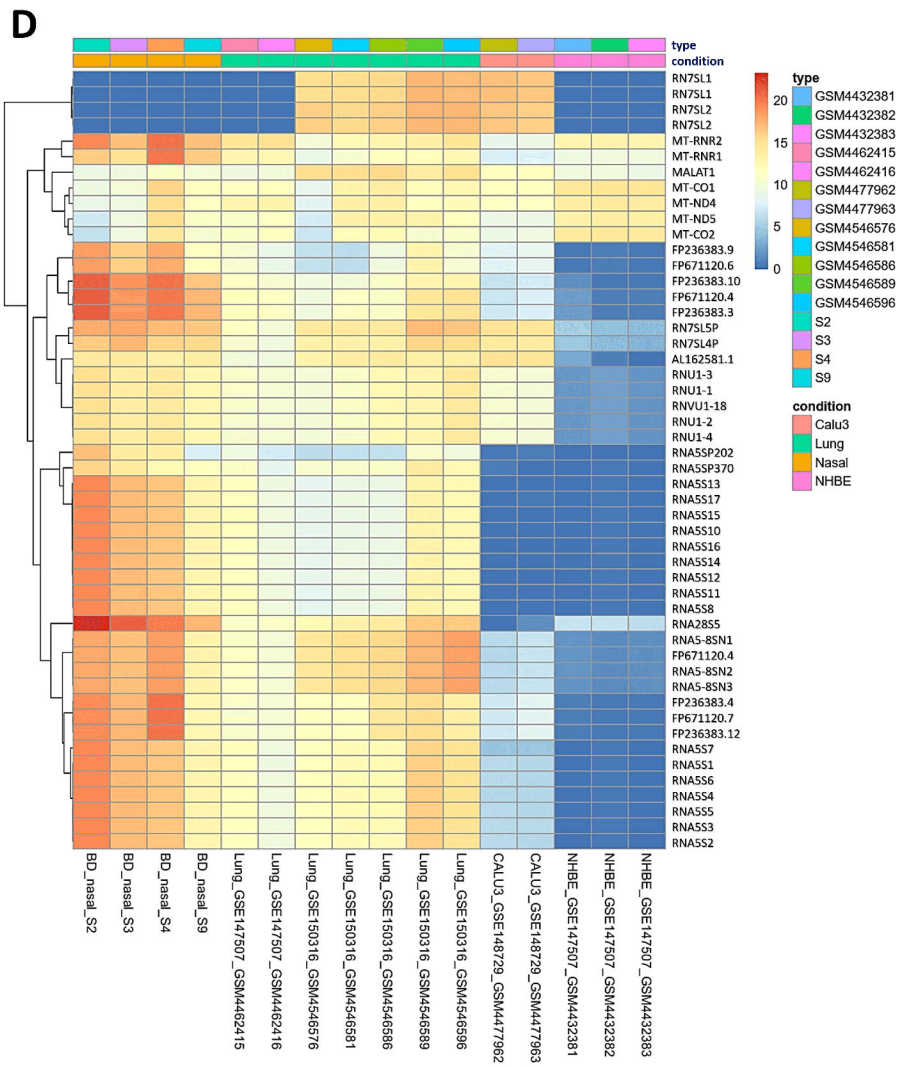


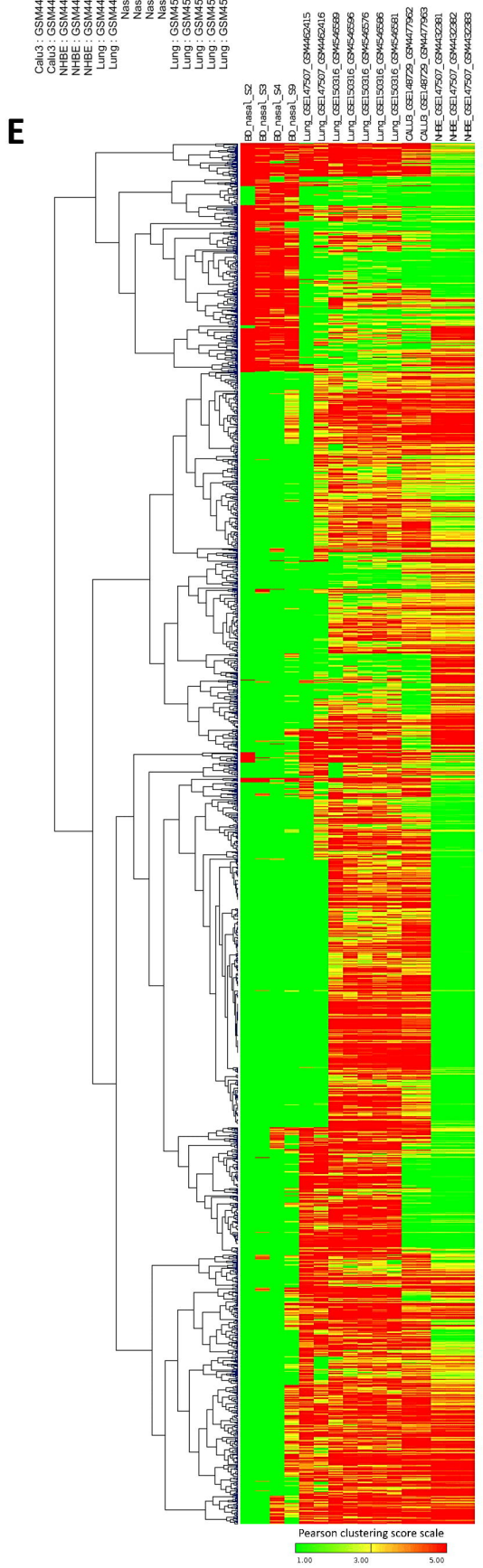
bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

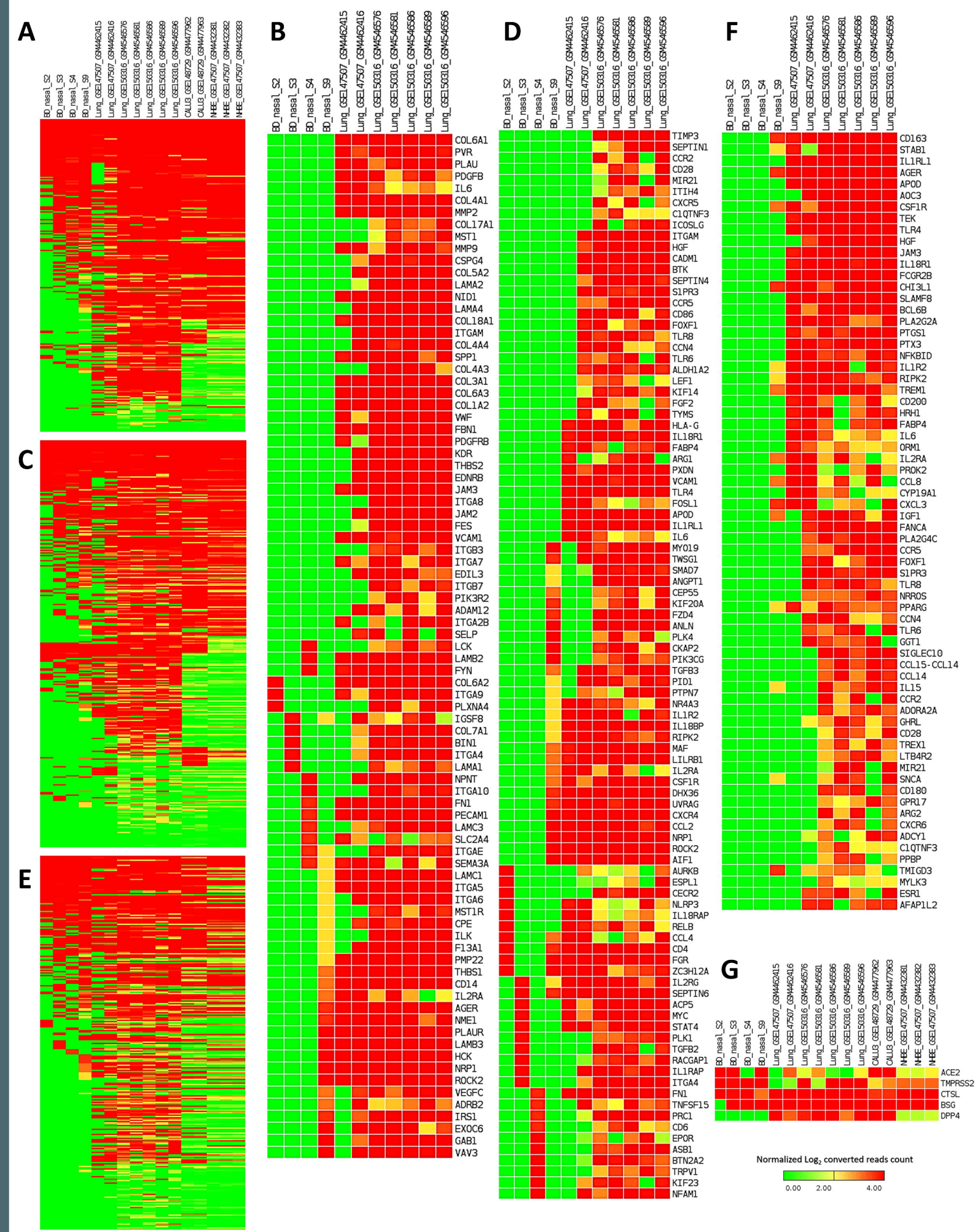


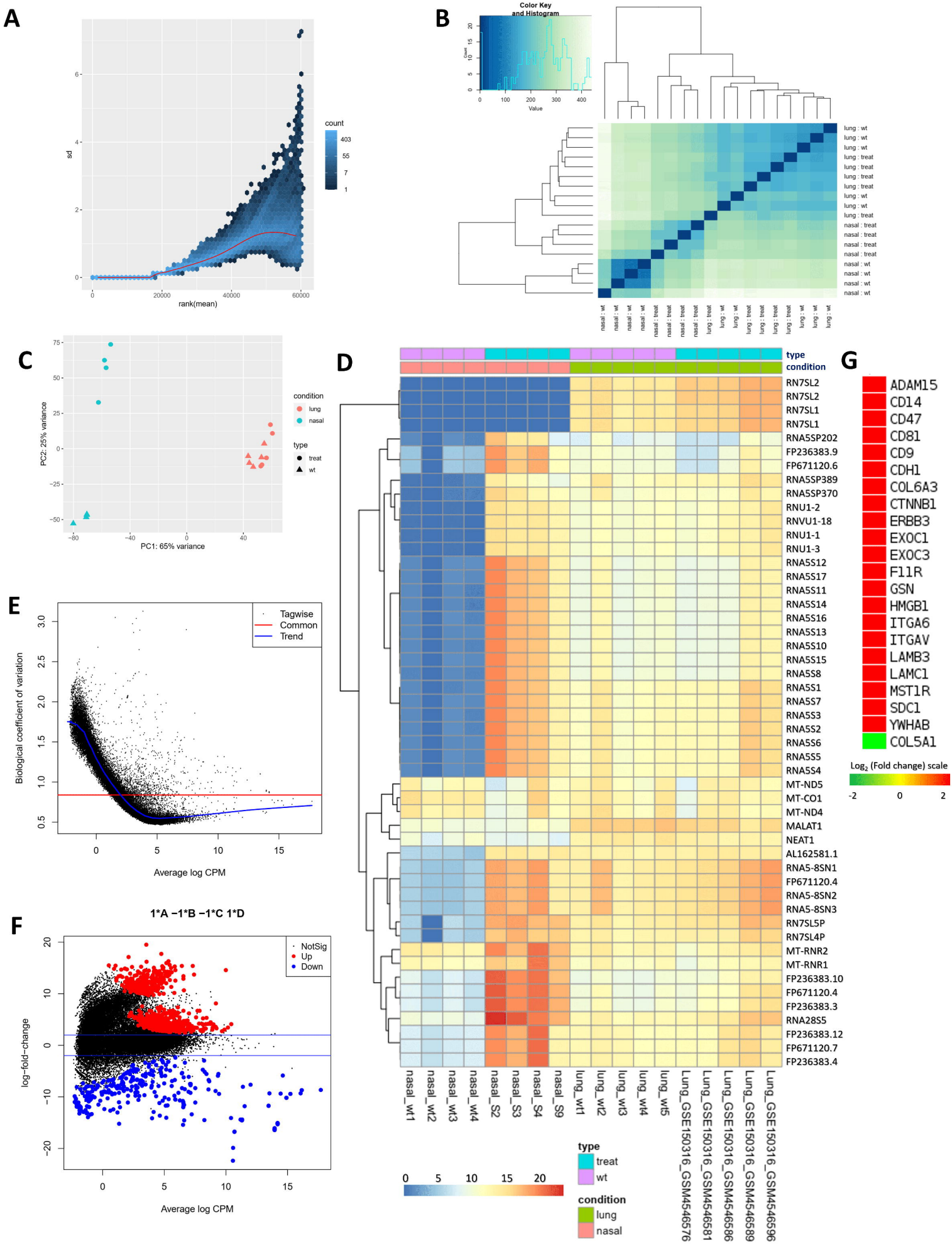




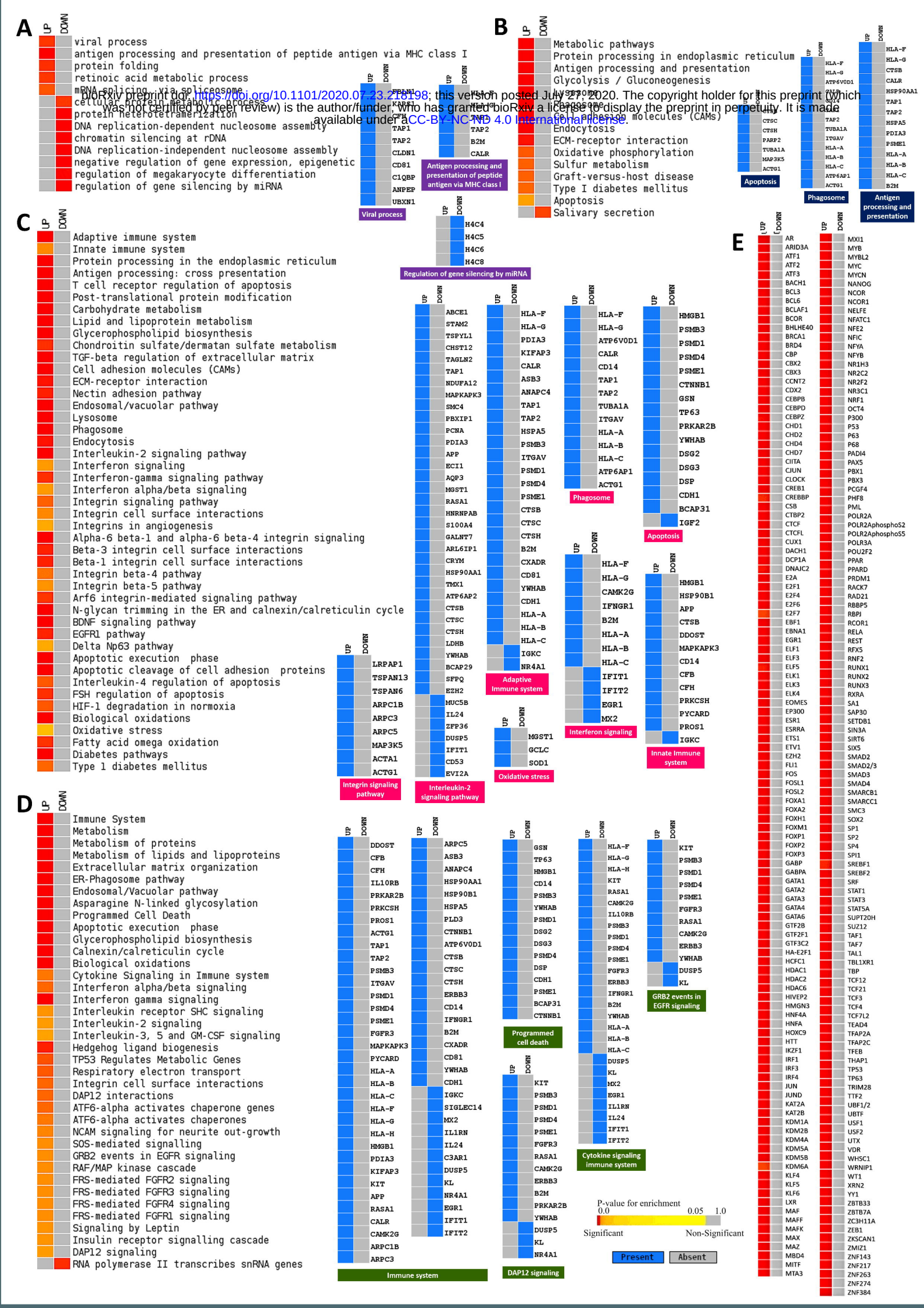




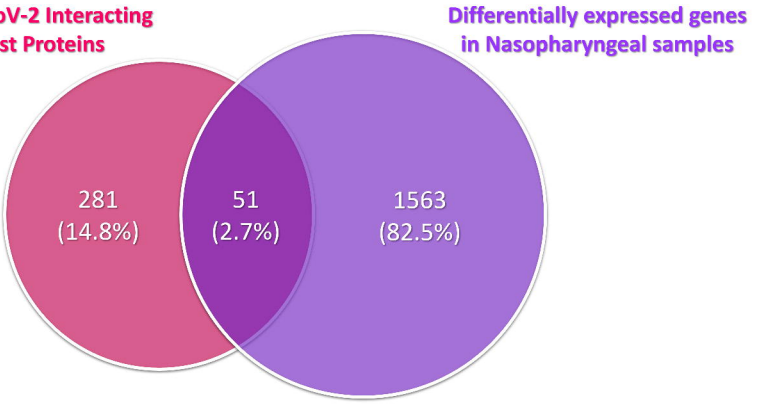

A
B
C
D
E
F
G
BD_nasal_S2
BD_nasal_S3
BD_nasal_S4
BD_nasal_S9
Lung_GSE147507_GSM4462415
Lung_GSE147507_GSM4462416
Lung_GSE150316_GSM4546576
Lung_GSE150316_GSM4546581
Lung_GSE150316_GSM4546586
Lung_GSE150316_GSM4546589
Lung_GSE150316_GSM4546596
CALU3_GSE148729_GSM4477962
CALU3_GSE148729_GSM4477963
NHBE_GSE147507_GSM4432381
NHBE_GSE147507_GSM4432382
NHBE_GSE147507_GSM4432383
COL6A1
PVR
PLAU
PDGFB
IL6
COL4A1
MMP2
COL17A1
MST1
MMP9
CSPG4
COL5A2
LAMA2
NID1
LAMA4
COL18A1
ITGAM
COL4A4
SPP1
COL4A3
COL3A1
COL6A3
COL1A2
VWF
FBN1
PDGFRB
KDR
THBS2
EDNRB
JAM3
ITGA8
JAM2
FES
VCAM1
ITGB3
ITGA7
EDIL3
ITGB7
PIK3R2
ADAM12
ITGA2B
SELP
LCK
LAMB2
FYN
COL6A2
ITGA9
PLXNA4
IGSF8
COL7A1
BIN1
ITGA4
LAMA1
NPNT
ITGA10
FN1
PECAM1
LAMC3
SLC2A4
ITGAE
SEMA3A
LAMC1
ITGA5
ITGA6
MST1R
CPE
ILK
F13A1
PMP22
THBS1
CD14
IL2RA
AGER
NME1
PLAUR
LAMB3
HCK
NRP1
ROCK2
VEGFC
ADRB2
IRS1
EXOC6
GAB1
VAV3
TIMP3
SEPTIN1
CCR2
CD28
MIR21
ITIH4
CXCR5
C1QTNF3
ICOSLG
ITGAM
HGF
CADM1
BTK
SEPTIN4
S1PR3
CCR5
CD86
FOXF1
TLR8
CCN4
TLR6
ALDH1A2
LEF1
KIF14
FGF2
TYMS
HLA-G
IL18R1
FABP4
ARG1
PXDN
VCAM1
TLR4
FOSL1
APOD
IL1RL1
IL6
MYO19
TWSG1
SMAD7
ANGPT1
CEP55
KIF20A
FZD4
ANLN
PLK4
CKAP2
PIK3CG
TGFB3
PID1
PTPN7
NR4A3
IL1R2
IL18BP
RIPK2
MAF
LILRB1
IL2RA
CSF1R
DHX36
UVRAG
CXCR4
CCL2
NRP1
ROCK2
AIF1
AURKB
ESPL1
CECR2
NLRP3
IL18RAP
RELB
CCL4
CD4
FGR
ZC3H12A
IL2RG
SEPTIN6
ACP5
MYC
STAT4
PLK1
TGFB2
RACGAP1
IL1RAP
ITGA4
FN1
TNFSF15
PRC1
CD6
EPOR
ASB1
BTN2A2
TRPV1
KIF23
NFAM1
CD163
STAB1
IL1RL1
AGER
APOD
AOC3
CSF1R
TEK
TLR4
HGF
JAM3
IL18R1
FCGR2B
CHI3L1
SLAMF8
BCL6B
PLA2G2A
PTGS1
PTX3
NFKBID
IL1R2
RIPK2
TREM1
CD200
HRH1
FABP4
IL6
ORM1
IL2RA
PROK2
CCL8
CYP19A1
CXCL3
IGF1
FANCA
PLA2G4C
CCR5
FOXF1
S1PR3
TLR8
NRROS
PPARG
CCN4
TLR6
GGT1
SIGLEC10
CCL15-CCL14
CCL14
IL15
CCR2
ADORA2A
GHRL
CD28
TREX1
LTB4R2
MIR21
SNCA
CD180
GPR17
ARG2
CXCR6
ADCY1
C1QTNF3
PPBP
TMIGD3
MYLK3
ESR1
AFAP1L2
ACE2
TMPRSS2
CTSL
BSG
DPP4
Normalized Log2 converted reads count
0.00
2.00
4.00


A
sd
rank(mean)
count
403
55
7
1
B
Color Key and Histogram
Value
Count
lung : wt
lung : treat
nasal : treat
nasal : wt
C
PC2: 25% variance
PC1: 65% variance
condition
lung
nasal
type
treat
wt
D
type
condition
RN7SL2
RN7SL1
RNA5SP202
FP236383.9
FP671120.6
RNA5SP389
RNA5SP370
RNU1-2
RNVU1-18
RNU1-1
RNU1-3
RNA5S12
RNA5S17
RNA5S11
RNA5S14
RNA5S16
RNA5S13
RNA5S10
RNA5S15
RNA5S8
RNA5S1
RNA5S7
RNA5S3
RNA5S2
RNA5S6
RNA5S5
RNA5S4
MT-ND5
MT-CO1
MT-ND4
MALAT1
NEAT1
AL162581.1
RNA5-8SN1
FP671120.4
RNA5-8SN2
RNA5-8SN3
RN7SL5P
RN7SL4P
MT-RNR2
MT-RNR1
FP236383.10
FP671120.4
FP236383.3
RNA28S5
FP236383.12
FP671120.7
FP236383.4
nasal_wt1
nasal_wt2
nasal_wt3
nasal_wt4
nasal_S2
nasal_S3
nasal_S4
nasal_S9
lung_wt1
lung_wt2
lung_wt3
lung_wt4
lung_wt5
Lung_GSE150316_GSM4546576
Lung_GSE150316_GSM4546581
Lung_GSE150316_GSM4546586
Lung_GSE150316_GSM4546589
Lung_GSE150316_GSM4546596
type
treat
wt
condition
lung
nasal
E
Tagwise
Common
Trend
Biological coefficient of variation
Average log CPM
F
1*A −1*B −1*C 1*D
NotSig
Up
Down
log−fold−change
Average log CPM
G
ADAM15
CD14
CD47
CD81
CD9
CDH1
COL6A3
CTNNB1
ERBB3
EXOC1
EXOC3
F11R
GSN
HMGB1
ITGA6
ITGAV
LAMB3
LAMC1
MST1R
SDC1
YWHAB
COL5A1
Log2 (Fold change) scale

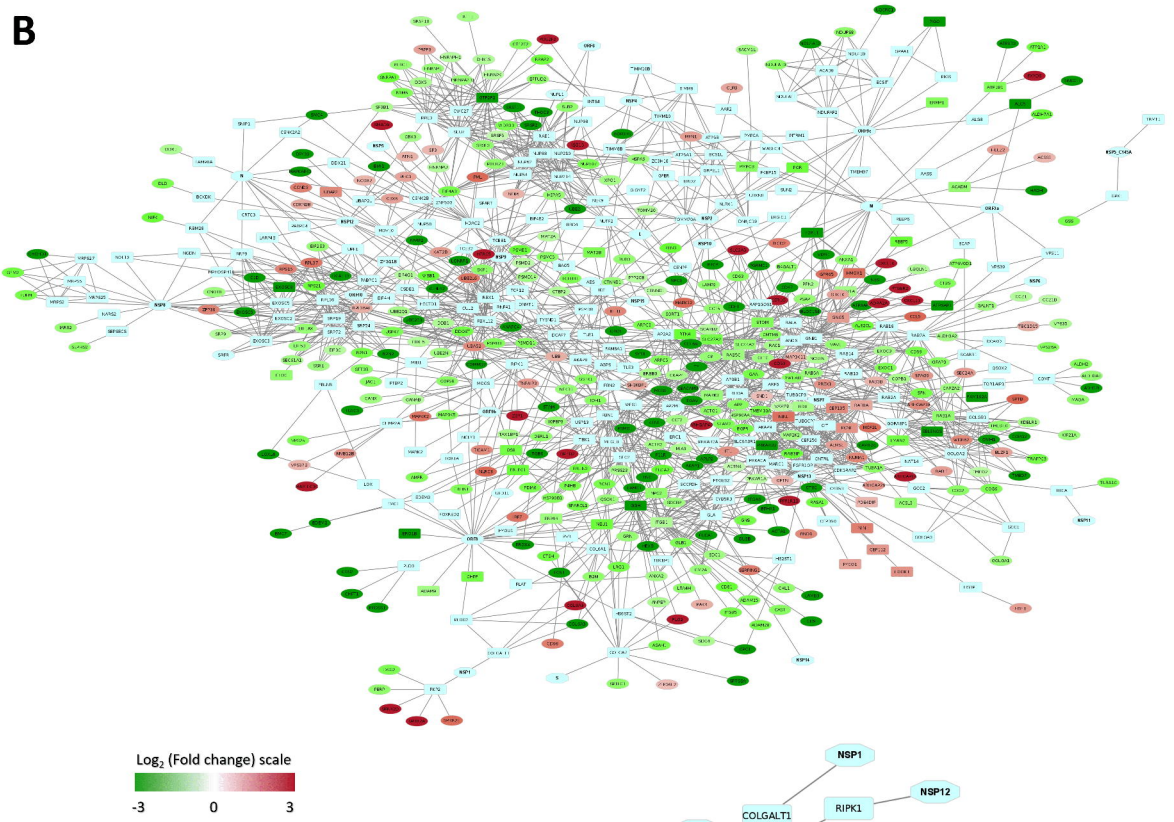
bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.218198; this version posted July 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



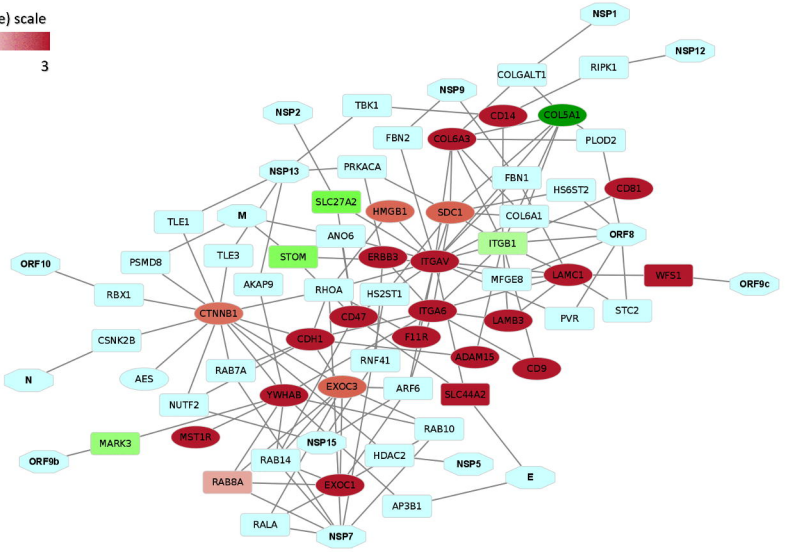
A


SARS-CoV-2 Interacting Host Proteins
Differentially expressed genes in Nasopharyngeal samples
281 (14.8%)
51 (2.7%)
1563 (82.5%)

B


Log2 (Fold change) scale
-3
0
3

C


NSP1
NSP12
COLGALT1
RIPK1
NSP9
TBK1
NSP2
CD14
COL5A1
FBN2
COL6A3
PLOD2
NSP13
PRKACA
FBN1
HS6ST2
CD81
SLC27A2
HMGB1
SDC1
COL6A1
TLE1
M
ORF8
ANO6
ITGB1
ORF10
PSMD8
TLE3
STOM
ERBB3
ITGAV
LAMC1
WFS1
ORF9c
AKAP9
RHOA
MFGE8
RBX1
HS2ST1
STC2
CTNNB1
CD47
ITGA6
LAMB3
PVR
CSNK2B
CDH1
F11R
ADAM15
CD9
N
AES
RAB7A
RNF41
EXOC3
ARF6
SLC44A2
YWHAB
NUTF2
RAB10
MARK3
MST1R
NSP15
ORF9b
RAB14
HDAC2
NSP5
E
RAB8A
EXOC1
AP3B1
RALA
NSP7