Heliyon 10 (2024) e32772

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Integrative genomic analysis of the lung tissue microenvironment in SARS-CoV-2 and NL63 patients

Krithika Bhuvaneshwar [*], Subha Madhavan, Yuriy Gusev [**]

Georgetown-Innovation Center for Biomedical Informatics (Georgetown-ICBI), Georgetown University Medical Center, Washington DC, 20007, USA

ARTICLE INFO

Keywords:
COVID19
SARS-COV-2
RNA-Seq
Viroinformatics
Genomics
Immunology
Metagenomics
Microbiome
Machine learning

ABSTRACT

The coronavirus disease 2019 (COVID-19) pandemic caused by the SARS-CoV-2 virus has affected over 700 million people, and caused over 7 million deaths throughout the world as of April 2024, and continues to affect people through seasonal waves. While over 675 million people have recovered from this disease globally, the lingering effects of the disease are still under study. Long term effects of SARS-CoV-2 infection, known as ’long COVID,’ include a wide range of symptoms including fatigue, chest pain, cellular damage, along with a strong innate immune response characterized by inflammatory cytokine production. Three years after the pandemic, data about long covid studies are finally emerging. More clinical studies and clinical trials are needed to understand and determine the factors that predispose individuals to these long-term side effects.

In this methodology paper, our goal was to apply data driven approaches in order to explore the multidimensional landscape of infected lung tissue microenvironment to better understand complex interactions between viral infection, immune response and the lung microbiome of patients with (a) SARS-CoV-2 virus and (b) NL63 coronavirus. The samples were analyzed with several machine learning tools allowing simultaneous detection and quantification of viral RNA amount at genome and gene level; human gene expression and fractions of major types of immune cells, as well as metagenomic analysis of bacterial and viral abundance. To contrast and compare specific viral response to SARS-COV-2, we analyzed deep sequencing data from additional cohort of patients infected with NL63 strain of corona virus.

Our correlation analysis of three types of RNA-seq based measurements in patients i.e. fraction of viral RNA (at genome and gene level), Human RNA (transcripts and gene level) and bacterial RNA (metagenomic analysis), showed significant correlation between viral load as well as level of specific viral gene expression with the fractions of immune cells present in lung lavage as well as with abundance of major fractions of lung microbiome in COVID-19 patients.

Our methodology-based proof-of-concept study has provided novel insights into complex regulatory signaling interactions and correlative patterns between the viral infection, inhibition of innate and adaptive immune response as well as microbiome landscape of the lung tissue. These initial findings could provide better understanding of the diverse dynamics of immune response and the side effects of the SARS-CoV-2 infection and demonstrates the possibilities of the various types of analyses that could be performed from this type of data.

[*] Corresponding author.

[**] Corresponding author.

E-mail addresses: kb472@georgetown.edu (K. Bhuvaneshwar), yg63@georgetown.edu (Y. Gusev).

https://doi.org/10.1016/j.heliyon.2024.e32772
Received 17 November 2023; Received in revised form 28 May 2024; Accepted 9 June 2024
Available online 10 June 2024

1. Introduction

The coronavirus disease 2019 (COVID-19) pandemic caused by the SARS-CoV-2 virus has affected over 700 million people and caused over 7 million deaths throughout the world as of April 2024, and continues to affect people through seasonal waves [1]. The majority of individuals infected were reported to have mild disease (about 75–80 %), about 15–20 % of patients need hospitalization, and about 5–10 % need critical care [2,3]. The estimated percentage of asymptomatic cases varied widely depending on factors such as population, age, frequency of testing, and virus strain. A meta-analysis study published in JAMA Network Open found that among 29 million individuals tested, asymptotic infections accounted for 0.25 % of the tested population and 40.50 % of the COVID positive population [4].

To date, a wide range of research articles on COVID-19 have been published including but not limited to the detection, treatment, prediction, and epidemiology of the disease. Researchers have also developed databases and curated data collections; and the large number of individuals affected lends itself to the application of various machine learning (ML) and artificial intelligence (AI) based algorithms in this domain. WHO had collated a research database that allows users to search through over 700, 000 articles related to COVID-19 published until Jan 2024 [5]. Other collections include Nature Communications' top 25 COVID-19 articles of 2022 and 2023, PMC COVID-19 collection and Springer Nature coronavirus collection [6–9]; and a few of the recent articles related to AI are also included here [10–12].

Although over 675 million people around the world have recovered from this disease, the long-term effects of the disease are still under investigation. Three years after the start of the pandemic, data from COVID-19 recovered patients showed multiple organs affected with a broad spectrum of manifestations. Long-term effects of SARS-CoV-2 infection, also known as 'long COVID' included fatigue, chest pain, cellular damage, and robust innate immune response with inflammatory cytokine production [2,13–15]. These collection of side effects are referred to as post-acute sequelae of SARS-CoV-2 infection (PASC) [16]. Large-scale efforts such as the NIH PASC Initiative [17] in the US and similar efforts in the UK [18] were launched with the goal to fund studies to build a biospecimen bank and track recovery. In 2023, NIH launched clinical trials for long COVID through the RECOVER Initiative to understand, treat, and prevent the condition. (https://recovercovid.org/).

According to a recent review article published in Nature, there is a conservative estimate of over 200 documented symptoms of Long COVID, and at least 65 million individuals around the world suffer from this disorder. Researchers have delved into the various risk factors, diagnosis, treatment, long-term health outcomes, and pathophysiological mechanisms which indicate likely multiple causes of long COVID. There is also a call to action for further research in order to improve health outcomes [19]. Despite this, there remains a notable gap in our understanding of the molecular landscape of Long COVID and the complex interactions within the infected microenvironments.

In this context, our methodology paper aims to address this gap by proposing a novel approach to the computational exploration of the multidimensional landscape of infected lung tissue microenvironment. Our goal is to better understand complex interactions between SARS-CoV-2 viral infection, immune response, and the lung microbiome of patients with SARS-CoV-2 or NL63 coronavirus. Our unique triple RNA-seq approach by exploring the interactions between pathogen, host, and microbiome has the potential to help researchers better understand the underlying mechanisms [20,21].

We would like to note that this work is not meant to be a validation study with large cohorts of patients. This computational exploration is meant to be a proof-of-concept article that demonstrates the possibilities of the various types of analyses that can be performed from this type of data.

In this article, we present the genomic analysis of deep sequencing data from publicly available RNA samples of lung lavage of COVID-19 patients. Utilizing three different machine learning tools, the analysis encompasses (a) simultaneous detection and quantification of viral RNA amount at genome and gene level; (b) human gene expression and fractions of major types of infiltrating immune cells, as well as (c) metagenomic analysis of bacterial and viral abundance in lower respiratory tract.

Furthermore, we compared and contrasted the specific immune responses to SARS-CoV-2 with the immune response to an infection by different coronavirus NL63. The NL63 coronavirus (HCoV-NL63) is known to cause severe lower respiratory tract infection, and bronchiolitis in vulnerable populations including children, the elderly and the immunocompromised [22]. To conduct this comparison, we applied the same computational tools and analyzed deep sequencing data from an additional cohort of patients infected with the NL63 strain of coronavirus.

We believe that our multi-dimensional exploratory study could provide novel insights into the complex regulatory signaling interactions and correlative patterns between the viral infection, inhibition of innate and adaptive immune response as well as microbiome landscape of the lung tissue. These initial findings could provide a better understanding of the diverse dynamics of immune response and the side effects of the SARS-CoV-2 infection.

2. Materials and methods

Our goal was to explore the multidimensional landscape of infected lung tissue microenvironment to better understand complex interactions between virus, immune response and microbiome in the lungs of COVID-19 patients. By utilizing three types of machine learning based bioinformatics tools, we were able to detect and quantitate three different fractions of short reads from RNA-seq data files: fraction of viral RNA (at genome and gene level), Human RNA (transcripts and gene level) and bacterial RNA (metagenomic analysis).

Our analysis pipeline is shown in Fig. 1. It includes three main sections: analysis and exploration of the viral RNA, human gene expression with a focus on immune cell expression signatures, and bacterial environment in the bulk RNA-seq data.

2.1. Datasets

For this paper, we applied our analysis workflow to two datasets described here. One of our goals was to explore sequences of the SARS-CoV-2 viral genome and gene level data and compare them to other coronavirus genomes. We chose a Human NL63 coronavirus (HCoV-NL63) dataset as this virus is known to affect children, the elderly and the immunocompromised [22]. This virus is one of the many human endemic coronaviruses (hCoV) that include HCoV-229E, and NL63-CoV that are from *alphacoronavirus* group; and OC43 and HKU1 that are from the *betacoronavirus* group [23]. This dataset allowed for direct comparison with the SARS-CoV-2 dataset and has been presented in the main body of the manuscript.

(a) ***SARS-CoV-2* Dataset:** We downloaded RNA-seq data from 5 patients affected with the SARS COV 2 virus from the NCBI SRA data repository PRJNA605983 (SRP249613) [24,25]. These 5 patients were from the early stage of the Wuhan seafood market pneumonia virus outbreak in China. The downloaded data were raw sequences in the form of *FASTQ* files. Total RNA was extracted from bronchoalveolar lavage fluid and then next generation sequencing (NGS) was performed using the Illumina platform. Out of the 5 samples, 4 were profiled on the Illumina HiSeq 3000 platform, and one on the Illumina HiSeq 1000 platform.

(b) NL63 coronavirus dataset (referred to as the *NL63-CoV* dataset in this paper):

We obtained a public dataset from 5 pediatric patients with severe lower respiratory infection by NL63 coronavirus with deep sequencing data performed on the Illumina HiSeq platform. The downloaded data were raw sequences in the form of *FASTQ* files obtained from NCBI PRJA601331 [26,27]. In this dataset, nasopharyngeal swab samples were obtained from children with severe acute respiratory infection (SIRS) and then sequenced using the Illumina HiSeq 2500 platform. The researchers who submitted the dataset noted in their manuscript (Zhang et al. [27]) that only partial genome sequences were obtained by NGS methods from the 5 samples [27].

2.2. Analysis of the viral environment in RNA-seq data

We first compiled our viral reference genome by downloading a combined dataset of viral sequences with humans as a host from the NCBI collection of viral genomes [28]. We then ran our viGEN viroinformatics pipeline [29] on the samples from the *SARS-CoV-2 and the NL63-CoV datasets*. viGEN is an open source bioinformatics pipeline that allows for not only the detection and quantification of viral RNA, but also the analysis of variants in the viral transcripts [29]. For this paper, we applied Bowtie2 aligner [30] in the viGEN pipeline to quantify the viral load detected in the two datasets. Bowtie2 is an alignment algorithm that uses a combination of index–assisted seed alignment and single-instruction multiple-data (SIMD) based dynamic parallel processing to achieve fast, accurate and sensitive alignment of sequencing data [30]. We also corroborated these results with the help of an additional pipeline CENTRIFUGE [31]. We

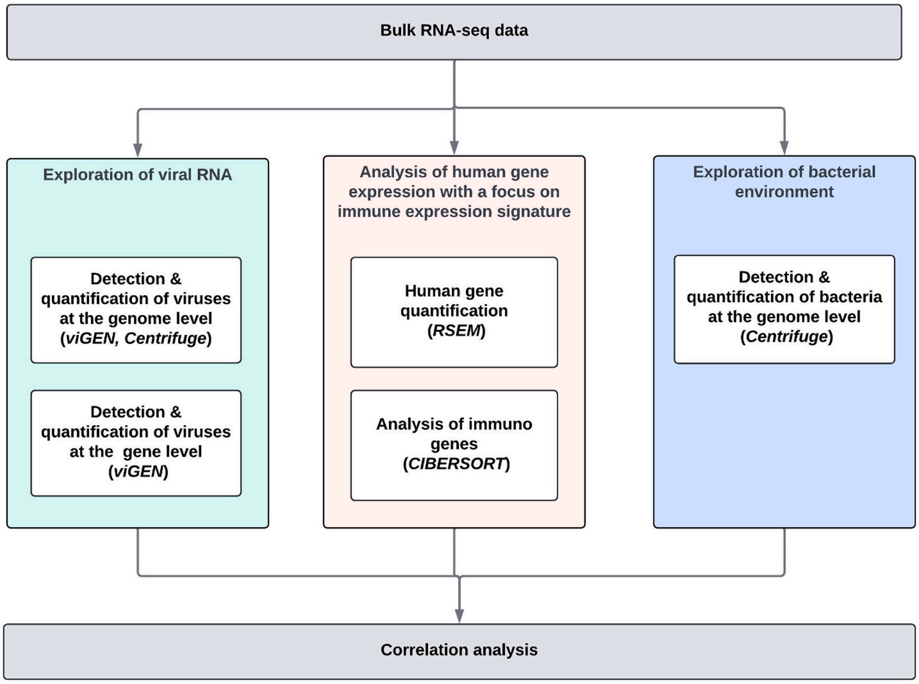

Bulk RNA-seq data
Exploration of viral RNA
Detection & quantification of viruses at the genome level (viGEN, Centrifuge)
Detection & quantification of viruses at the gene level (viGEN)
Analysis of human gene expression with a focus on immune expression signature
Human gene quantification (RSEM)
Analysis of immuno genes (CIBERSORT)
Exploration of bacterial environment
Detection & quantification of bacteria at the genome level (Centrifuge)
Correlation analysis

Fig. 1. Data analysis workflow.

also applied our quantification algorithm of viral RNA at the gene/CDS level, which is part of the viGEN pipeline [29]. This produced gene counts of viral RNA on the input datasets.

2.3. Analysis of the human immune environment in the RNA-seq data based on gene expression data

The raw sequences from the *SARS-COV-2 and NL63-CoV* RNA-seq datasets were first aligned to the human reference genome version hg38 using an open source RNA-seq alignment and quantification pipeline that used the RSEM-STAR aligner algorithm [32]. This analysis was performed in the Seven Bridges (SB) Cancer Genomics Cloud (CGC) Platform [33,34]. The output of this pipeline was gene and isoform level data in the form of both raw counts and transcripts per million (TPM) values. Out of the ~20,000 genes in the human genome, the gene quantification data in the form of TPM values were extracted for a subset of 530 immune related genes.

We then applied our immuno-genomics pipeline to the gene quantification output. This gene matrix was input into a public online tool CIBERSORT [35]. CIBERSORT is a virtual flow cytometry tool that estimates the abundance of immune cell types in the samples using gene expression from microarrays or RNA-seq data. It is a machine learning algorithm that uses nu-linear support vector regression (ν-SVR) to perform deconvolution of the input mixture. N-SVR is a type of support vector machine (SVM) wherein a hyperplane that maximally separates both classes is discovered [36]. The CIBERSORT analysis was performed with quantile normalization disabled and permuted 500 times. CIBERSORT uses a signature matrix of known immune cell mixtures to estimate the immune cell fractions of the input dataset. We used the LM22 signature matrix which was built by the software creators using a set of 547 genes that could accurately distinguish 22 mature human hematopoietic populations isolated from peripheral blood or in vitro culture conditions, including seven T cell types, naïve and memory B cells, plasma cells, NK cells, and myeloid subsets. The output of CIBERSORT was in the form of estimated fractions of 22 immune cell types across the input samples [36].

2.4. Analysis of the bacterial environment in the RNA-seq data

Additional analysis was performed on the same RNA samples by applying the metagenomics pipeline CENTRIFUGE [31] to detect and quantitate the abundance of bacterial species and viruses comprising the lung microbiome. Centrifuge is a machine learning algorithm that uses an indexing scheme based on the Burrows-Wheeler transform (BWT) and the Ferragina-Manzini (FM) index. It was designed and optimized specifically for metagenomics applications [31]. We ran the Centrifuge analysis pipeline on the Seven Bridges CGC platform [33,34] on the input same FASTQ files from the datasets. The index file containing the reference genomes for human, prokaryotic genomes, and viral genomes including 106 SARS-CoV-2 complete genomes was provided by the CENTRIFUGE team on their website [37]. We also used the online web application Pavian [38,39] for visual analysis of results generated with Centrifuge.

2.5. Correlation analysis

Once the three types of results were obtained for each sample in the datasets, downstream correlation analysis was performed. This allowed us to explore possible interactions and regulation of viral infection with the local immunological environment in the lungs of infected patients, as well as, microbiome profiles measured by changes in the abundance of multiple bacterial species in the lung. The Pearson correlation analysis was performed in the R statistical software [40].

3. Results

In this proof-of-concept paper, our goal was to explore the multidimensional landscape of infected lung tissue microenvironment to better understand complex interactions between virus, immune response and microbiome in the lungs of COVID-19 patients. By utilizing three types of bioinformatics workflows and tools, we were able to detect and quantitate three different fractions of short reads from RNA-seq data files: fraction of viral RNA (at genome and gene level), Human RNA (transcripts and gene level) and bacterial RNA (metagenomic analysis). Correlation analysis of these three types of measurements in patients has shown significant correlation between viral load as well as the level of specific viral gene expression with the fractions of immune cells present in lung lavage as well as with the abundance of major fractions of the lung microbiome.

3.1. Section 1: Results of analysis on the SARS-CoV-2 dataset

3.1.1. Results of analysis of the viral environment

Supplementary File 1A shows the estimated copy number of viral genomes detected in lung lavage samples of the *SARS-CoV-2 dataset* obtained using the viGEN pipeline. We can clearly see the presence of the various strains SARS-CoV-2 virus in all of the 5 patients in this dataset. The SARS-CoV-2 virus in these samples has also been detected and corroborated using the CENTRIFUGE metagenomics pipeline (Supplementary File 1B).

Supplementary File 1C shows the estimated level of viral gene expression counts in the patients from the *SARS-CoV-2 dataset*. The *three prime UTR* regions in the region ranging from nucleotide position 29675 to 29903 showed one of the largest abundances of viral gene expression with raw gene counts of more than 1000 for most patients. We also see the presence of the older SARS coronavirus (SARS-CoV) which was prevalent in Asia in 2003 [41]. This makes sense as these RNA-seq samples were obtained from patients in China. As expected, we see the presence of flu and common cold viruses including Human Coronavirus 229E and Influenza A virus (A/Shanghai/02/2013(H7N9)).

3.1.2. Results of analysis of the human immune environment

Next, we assessed the immuno-profile of samples from the *SARS-CoV-2 dataset* (Fig. 2A). It shows a summary of the estimated fractions of 22 types of immune cells detected in lung lavage and indicates Macrophages M2, T cells CD4 naïve, Natural Killer (NK) cells resting and Monocytes as the most abundant types of immune cells in this dataset.

3.1.3. Results of analysis of the bacterial environment

The metagenomic analysis of the patients in the *SARS-CoV-2 dataset* (Table 1) shows the abundance of the top 20 bacterial species in the lung microbiome. Fig. 3 shows a Sankey diagram visualization of the bacterial species. A Sankey diagram [42] is a flow diagram in which the width of the arrows is proportional to the abundance (i.e. number of reads) of the bacterial species, which are collated together by the taxonomy of bacteria. Hence the most abundant bacterial species will be shown on the right-most side with the largest width. The Sankey diagrams indicated bacterial species *Clostridium botulinum* and *Clostridium tetani* to be one of the most abundant species in these samples.

3.1.4. Results of correlation between genomic and immunological data

We found a significant correlation between viral load (genomic and gene level) and the immune-profile of the patients in the *SARS-*

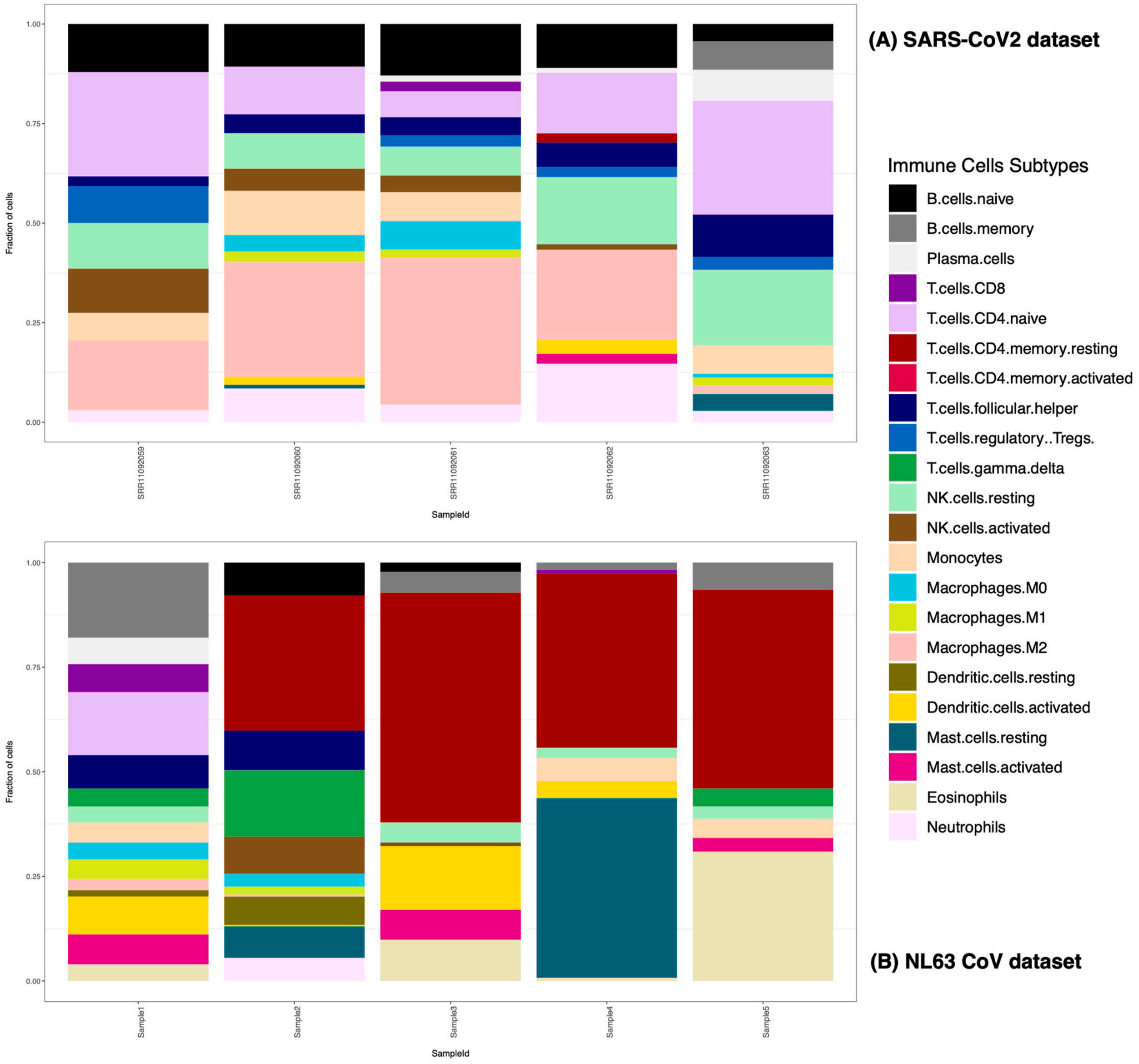


Fig. 2. CIBERSORT output: estimated fractions of 22 types of immune cells detected in lung lavage of (A) 5 patients from the SARS-COV-2 dataset and (B) 5 patients from the NL63-COV-Hiseq dataset. There is one stacked bar per patient.

Table 1
Abundance of top bacterial species in lung microbiome of patients in the SARS-CoV-2 dataset. Showing top 20 sorted based on minimum counts.

Name	Minimum	SRR11092059	SRR11092060	SRR11092061	SRR11092062	SRR11092063
Clostridium tetani	257477	886195	305193	257477	433532	355297
Clostridium botulinum	218725	1499933	252669	218725	317989	264103
Trichormus azollae	29537	31171	37568	29537	57222	46985
Spirosoma pollinicola	19190	246664	24916	25987	24161	19190
Prevotella oris	9083	9757	9986	9083	13242	10862
Staphylococcus aureus	8482	1142840	46143	28001	13914	8482
Stenotrophomonas maltophilia group	7784	1312705	116353	116221	24795	7784
Stenotrophomonas maltophilia	7784	1312705	116353	116221	24795	7784
Prevotella denticola	7630	7955	8340	7630	10548	8846
Prevotella fusca	7036	7530	7895	7036	10089	8459
Roseimicrobium sp. ORNL1	6499	6872	8053	6499	12297	10272
Enterobacter cloacae complex	6377	698959	68035	67773	13735	6377
Prevotella ruminicola	5829	6073	6150	5829	8186	6958
Enterobacter kobei	4130	452826	44783	44580	8897	4130
Pseudomonas putida group	4105	556693	51160	51556	11347	4105
Pseudomonas putida	3667	516644	47579	47876	10105	3667
Sphingomonas paucimobilis	1766	296679	1766	23579	13965	2448
Enterobacter cloacae	1461	160877	15373	15426	3054	1461
Lacrimispora saccharolytica	1140	159024	13606	13991	2903	1140
Total Bacterial reads in the sample	**984975**	**10726961**	**1358144**	**1328864**	**1353674**	**984975**
Total Microbial reads in the sample	**1016348**	**10849767**	**1412506**	**1578271**	**1436901**	**1016348**

CoV-2 dataset.

Fig. 4 (A and B) shows the summary of the statistically significant correlations between viral gene expression and immunological cell types for the SARS-CoV-2 dataset. The full correlation matrix is provided in Supplementary File 2A (genome level) and Supplementary File 2B (gene level). Fig. 4A shows the statistically significant correlation results between viral load (genome level) and fraction of immune cells. Results show that Macrophages M2 are positively correlated with the SARS-CoV-2 virus genome counts, while NK cells activated, monocytes and T cells CD4 Naïve were negatively correlated with genome counts. Fig. 4B represents statistically significant correlations between viral gene expression (gene level) and fraction of immune cells. Activated NK cells and monocytes were found to be inversely correlated with the gene counts of the 3 prime and 5 prime UTR regions respectively. Monocytes were also found to be inversely correlated with other regions of the SARS-CoV-2 virus genome including membrane glycoprotein, envelope protein, nucleocapsid phosphoprotein and more.

3.1.5. *Results of correlation between bacterial abundance with immunological cell types*

We chose a p-value cut off of 0.005 to get a short list of 70 statistically significant correlation results. Out of these 70 short listed results, only one was negatively correlated, while the rest of the 69 results were positively correlated. Monocytes were negatively correlated with *Schaalia odontolytica* bacterial species. NK cells activated and Eosinophils were positively correlated with the following families including *Citrobacter, Clostridium, Delftia, Enterobacter, Lactobacillus, Paenibacillus, Phyllobacterium and Pseudomonas, Staphylococcus and Stenotrophomonas* (Fig. 5A). The complete correlation results for this correlation analysis are available as Supplementary File 2C.

3.1.6. *Results of correlation between viral load (genomic level) with bacterial abundance*

We chose a p-value cut off of 0.005 to get a short list of 261 statistically significant features that were correlated between viral load (genomic level) and bacterial abundance. Due to the large number of results, we focused on the correlation results in the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) genome. The complete correlation results for this correlation analysis are available as Supplementary File 2D. Of the 261 statistically significant features, only one bacterial species *Schaalia odontolytica* was positively correlated with the SARS-CoV-2 genome. The rest of the results were negatively correlated with the SARS-CoV-2 genome including *Citrobacter, Enterobacter, Lactobacillus, Paenibacillus, and Pseudomonas*. These results are summarized in Fig. 5B).

3.2. *Section 2: Results of analyses on the NL63-CoV dataset*

3.2.1. *Results of analysis of the viral environment*

Supplementary File 3A shows the estimated viral genome copy numbers in 5 pediatric patients from the *NL63-CoV dataset* obtained using the viGEN pipeline. Human coronavirus NL63 virus was detected in only two of the five pediatric patients. This same result is corroborated using the Centrifuge pipeline (Supplementary File 3B) which showed a read count of more than 100 for the same two samples. The researchers who submitted the dataset noted in their manuscript (Zhang et al. [27]) that only partial genome sequences were obtained by NGS methods from the 5 samples [27]. We think this may have caused the non-detection of the NL63-CoV virus in some samples.

Supplementary File 3C shows the NL63 viral gene/CDS counts in the *NL63-CoV dataset.* Some of the most abundant regions

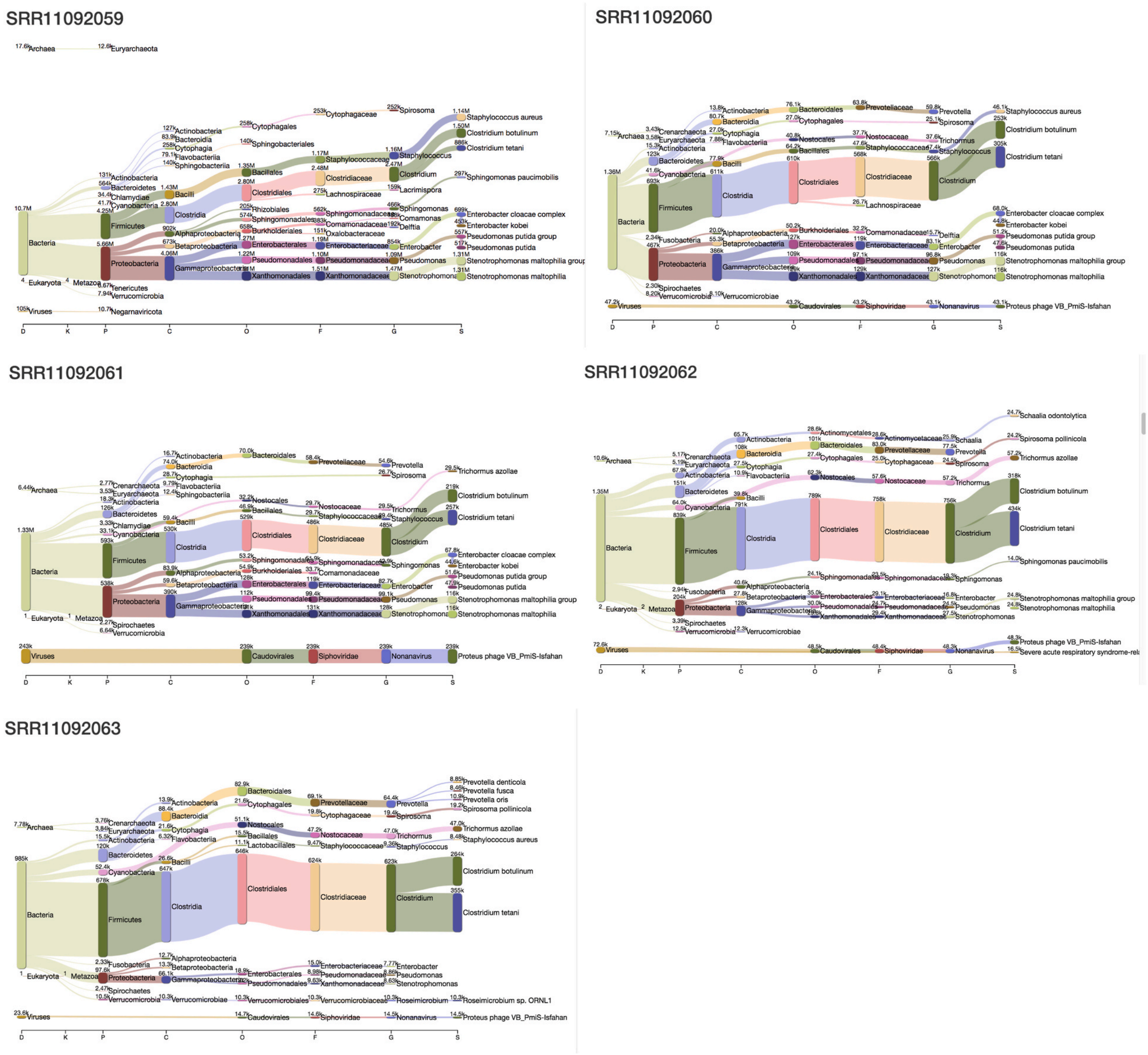


Fig. 3. Lung microbiome profile for the SARS-CoV-2 dataset represented as a Sankey diagram visualization of the bacterial species.

included the CDS regions in the HCNV63gp1 and HCNV63gp2 genes that code for the *replicase polyprotein 1 ab*. Other dominant regions include a CDS region in the HCNV63gp2 gene that codes for a spike protein; and a CDS region in the HCNV63gp4 gene that codes for an envelope protein (small membrane protein).

3.2.2. *Results of analysis of the immune environment*

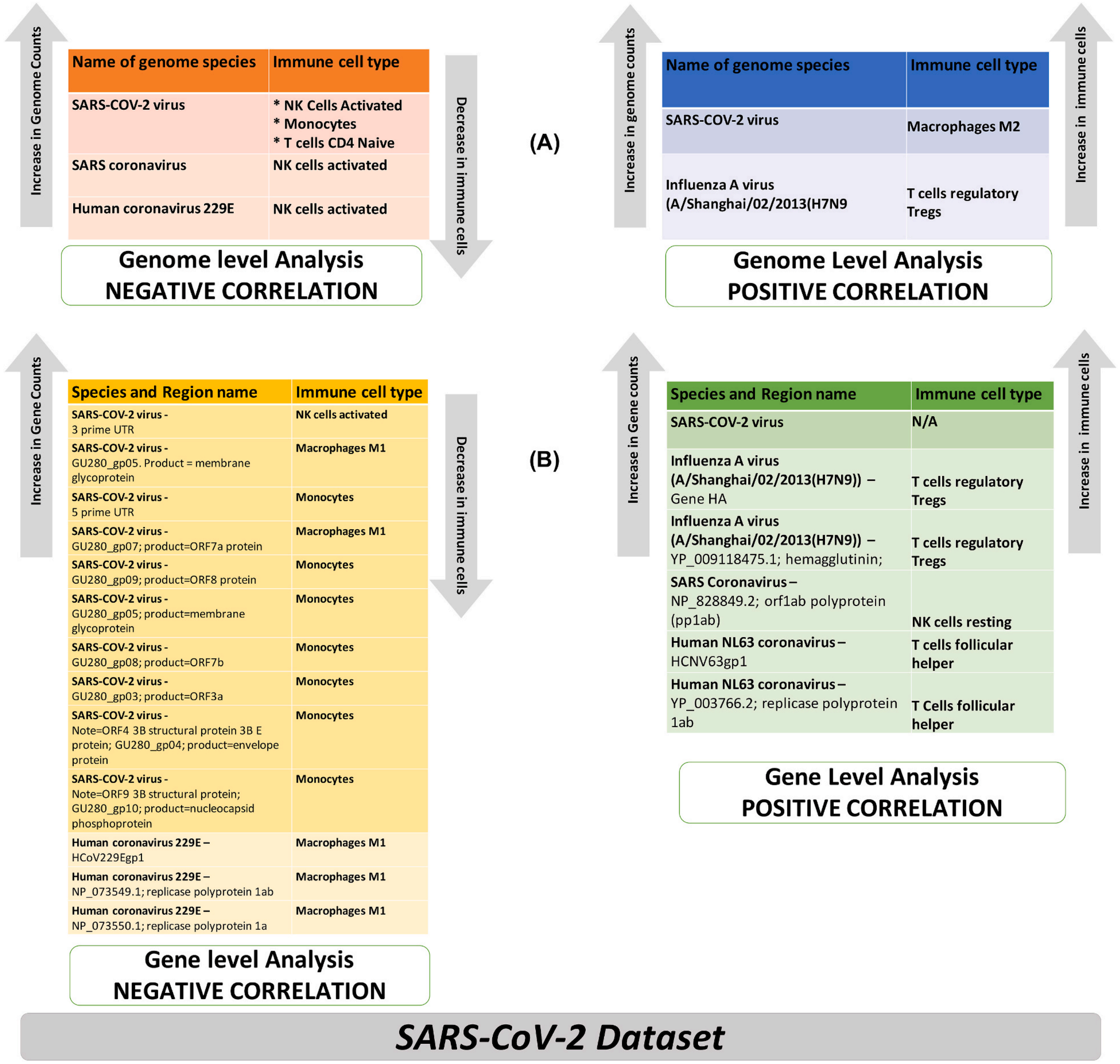
Next, we assessed the immuno-profile of samples from the *NL63-CoV dataset* (Fig. 2B). It indicates T cells CD4 memory as the dominant immune cells.

3.2.3. *Results of analysis of the bacterial environment*

The metagenomic analysis of the patients in the *NL63-CoV dataset* (Table 2) shows the abundance of the top 20 bacterial species in the nasopharyngeal microenvironment showing *Streptococcus pneumoniae* as one of the dominant species. Supplementary File 4 shows the nasopharyngeal microbiome profile of pediatric patients from the *NL63-CoV dataset* represented as a Sankey diagram visualization of the bacterial species.

3.2.4. *Results of correlation between genomic and immunological data*

We did not find many significant correlations between viral load and viral gene expression and the immune-profile of the patients in



(A)

Increase in Genome Counts

Decrease in immune cells

Name of genome species	Immune cell type
SARS-COV-2 virus	* NK Cells Activated * Monocytes * T cells CD4 Naive
SARS coronavirus	NK cells activated
Human coronavirus 229E	NK cells activated

Genome level Analysis
NEGATIVE CORRELATION

Increase in genome counts

Increase in immune cells

Name of genome species	Immune cell type
SARS-COV-2 virus	Macrophages M2
Influenza A virus (A/Shanghai/02/2013(H7N9	T cells regulatory Tregs

Genome Level Analysis
POSITIVE CORRELATION

(B)

Increase in Gene Counts

Decrease in immune cells

Species and Region name	Immune cell type
SARS-COV-2 virus - 3 prime UTR	NK cells activated
SARS-COV-2 virus - GU280_gp05. Product = membrane glycoprotein	Macrophages M1
SARS-COV-2 virus - 5 prime UTR	Monocytes
SARS-COV-2 virus - GU280_gp07; product=ORF7a protein	Macrophages M1
SARS-COV-2 virus - GU280_gp09; product=ORF8 protein	Monocytes
SARS-COV-2 virus - GU280_gp05; product=membrane glycoprotein	Monocytes
SARS-COV-2 virus - GU280_gp08; product=ORF7b	Monocytes
SARS-COV-2 virus - GU280_gp03; product=ORF3a	Monocytes
SARS-COV-2 virus - Note=ORF4 3B structural protein 3B E protein; GU280_gp04; product=envelope protein	Monocytes
SARS-COV-2 virus - Note=ORF9 3B structural protein; GU280_gp10; product=nucleocapsid phosphoprotein	Monocytes
Human coronavirus 229E – HCoV229Egp1	Macrophages M1
Human coronavirus 229E – NP_073549.1; replicase polyprotein 1ab	Macrophages M1
Human coronavirus 229E – NP_073550.1; replicase polyprotein 1a	Macrophages M1

Gene level Analysis
NEGATIVE CORRELATION

Increase in Gene counts

Increase in immune cells

Species and Region name	Immune cell type
SARS-COV-2 virus	N/A
Influenza A virus (A/Shanghai/02/2013(H7N9)) – Gene HA	T cells regulatory Tregs
Influenza A virus (A/Shanghai/02/2013(H7N9)) – YP_009118475.1; hemagglutinin;	T cells regulatory Tregs
SARS Coronavirus – NP_828849.2; orf1ab polyprotein (pp1ab)	NK cells resting
Human NL63 coronavirus – HCNV63gp1	T cells follicular helper
Human NL63 coronavirus – YP_003766.2; replicase polyprotein 1ab	T Cells follicular helper

Gene Level Analysis
POSITIVE CORRELATION

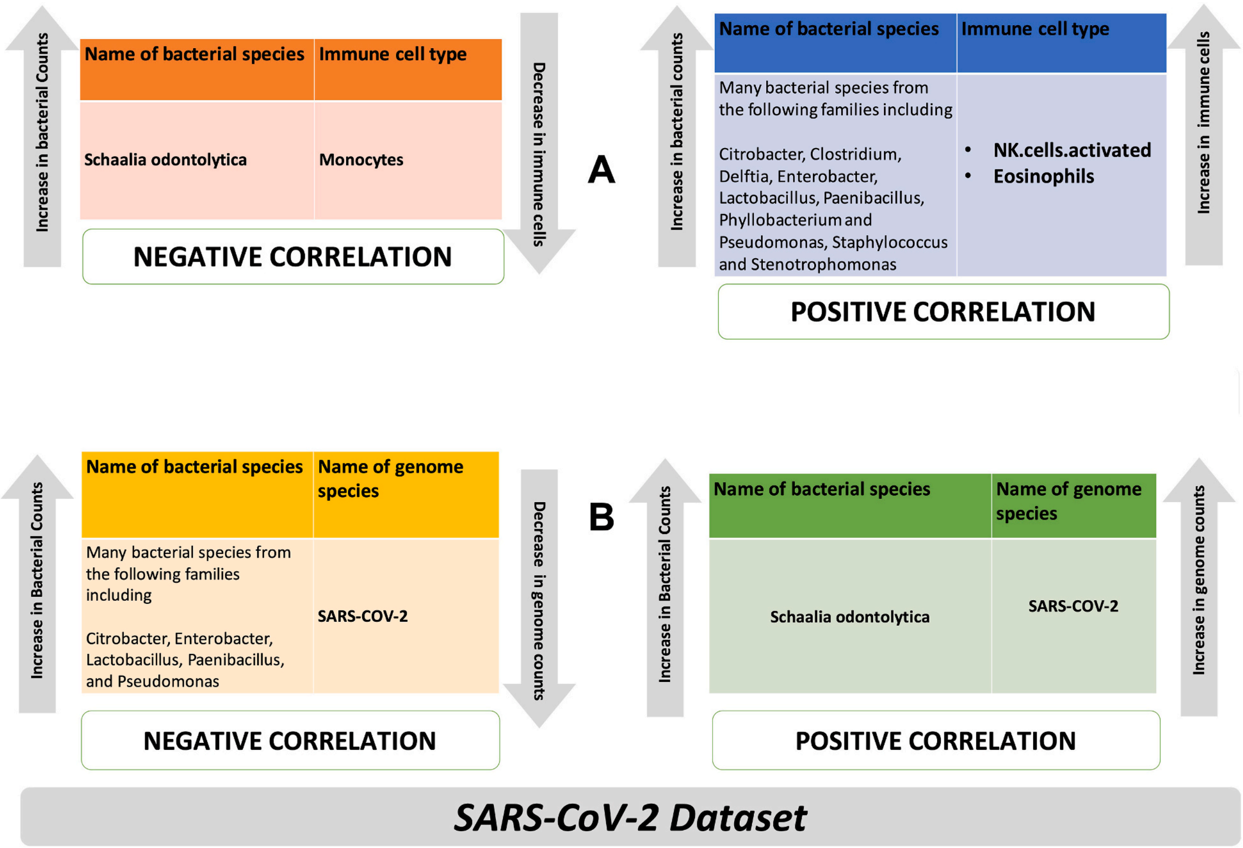
SARS-CoV-2 Dataset

Fig. 4. Summary of the statistically significant correlations between viral gene expression and immunological cell types for the SARS-CoV-2 dataset. (A) Viral genome level correlation between. viral load and fraction of immune cells (B). Viral gene level correlation between viral gene expression and fraction of immune cells.

the *NL63-CoV dataset.* This may have been attributed to the challenges the owners of this dataset faced with regard to the partial genome sequences obtained during sequencing by NGS methods. While there was no significant correlation between viral load and viral gene expression and immune-profile for the NL63 coronavirus species, we did find some significant correlation with another similar coronavirus species, the Human Coronavirus 229E. Fig. 7 shows the summary of the statistically significant correlations between viral gene expression and immunological cell types for the *NL63-CoV dataset*. The full correlation matrix is provided as Supplementary File 5A (genome level) and Supplementary File 5B (gene level). Fig. 6A represents the correlation at the genome level between viral load and fraction of immune cells. Fig. 6B represents the gene level correlation between viral gene expression and estimated fractions of immune cells.

3.2.5. Correlation between metagenomic bacterial abundance with immunological cell types

We chose a p-value cut off of 0.005 to get a short list of 17 statistically significant correlation results. Out of these 17 short listed results, 16 were negatively correlated, and one was positively correlated. B cells naïve were positively correlated with bacterial species *Mycoplasma orale.* Monocytes and Mast Cells resting were negatively correlated with many bacterial species from the following families



A

Increase in bacterial Counts

Name of bacterial species	Immune cell type
Schaalia odontolytica	Monocytes

Decrease in immune cells

NEGATIVE CORRELATION

Increase in bacterial counts

Name of bacterial species	Immune cell type
Many bacterial species from the following families including Citrobacter, Clostridium, Delftia, Enterobacter, Lactobacillus, Paenibacillus, Phyllobacterium and Pseudomonas, Staphylococcus and Stenotrophomonas	• NK.cells.activated • Eosinophils

Increase in immune cells

POSITIVE CORRELATION

B

Increase in Bacterial Counts

Name of bacterial species	Name of genome species
Many bacterial species from the following families including Citrobacter, Enterobacter, Lactobacillus, Paenibacillus, and Pseudomonas	SARS-COV-2

Decrease in genome counts

NEGATIVE CORRELATION

Increase in Bacterial Counts

Name of bacterial species	Name of genome species
Schaalia odontolytica	SARS-COV-2

Increase in genome counts

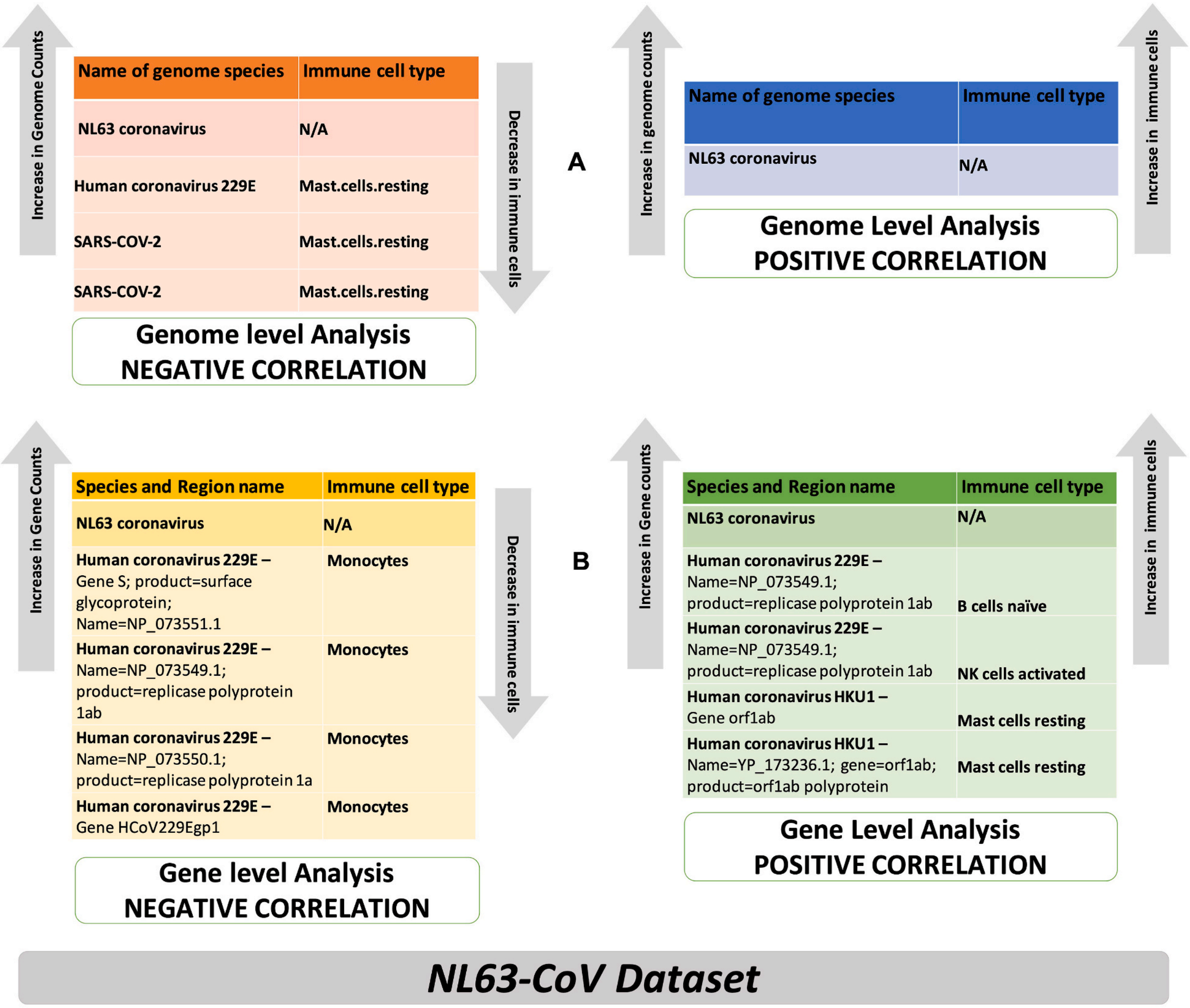
POSITIVE CORRELATION

SARS-CoV-2 Dataset

Fig. 5. (A): Summary of correlation analysis between bacterial abundance with immunological cell types for the SARS-CoV-2 dataset (B): Correlation analysis between viral load (genomic level) with bacterial abundance in the SARS-CoV-2 dataset.

Table 2
List of top bacterial species dominating nasopharyngeal microenvironment microbiome of 5 NL63 patients based on maximum abundance. *Showing top 20 sorted based on minimum counts.*

Name	Minimum	Max	Sample1	Sample2	Sample3	Sample4	Sample5
Streptococcus pneumoniae	5121	196870	14355	39493	196870	5121	12381
Streptococcus mitis	4742	465843	28008	74808	465843	4742	20932
Campylobacter concisus	3187	536664	12525	536664	50631	3255	3187
Prevotella melaninogenica	2604	1058483	1058483	587266	395090	2604	13004
Haemophilus influenzae	2408	33233	10396	15293	33233	2408	23473
Streptococcus pseudopneumoniae	2265	72017	4201	9186	72017	2265	4597
Prevotella jejuni	1770	818696	41903	552126	818696	2692	1770
Haemophilus haemolyticus	1710	44072	14249	23371	31369	1710	44072
Veillonella dispar	1709	360869	54288	360869	347922	1709	3706
Streptococcus salivarius	1381	103093	1650	33522	103093	1381	3776
Veillonella atypica	1169	470491	132008	470491	392930	1878	1169
Neisseria meningitidis	1083	23338	12358	23338	4311	1083	2798
Neisseria flavescens	908	293417	293417	46429	12337	1595	908
Streptococcus gwangjuense	805	83898	5497	17502	83898	805	3638
Streptococcus sp. 116-D4	637	48931	3913	16995	48931	637	2940
Haemophilus sp. oral taxon 036	610	9971	2992	4245	9971	610	8280
Fusobacterium pseudoperiodonticum	530	3451701	3451701	1120955	27430	530	596
Prevotella intermedia	513	268641	165720	268641	99410	513	1128
Streptococcus oralis	502	82478	11027	81068	82478	502	2354



Name of genome species	Immune cell type
NL63 coronavirus	N/A
Human coronavirus 229E	Mast.cells.resting
SARS-COV-2	Mast.cells.resting
SARS-COV-2	Mast.cells.resting

Genome level Analysis
NEGATIVE CORRELATION

Name of genome species	Immune cell type
NL63 coronavirus	N/A

Genome Level Analysis
POSITIVE CORRELATION

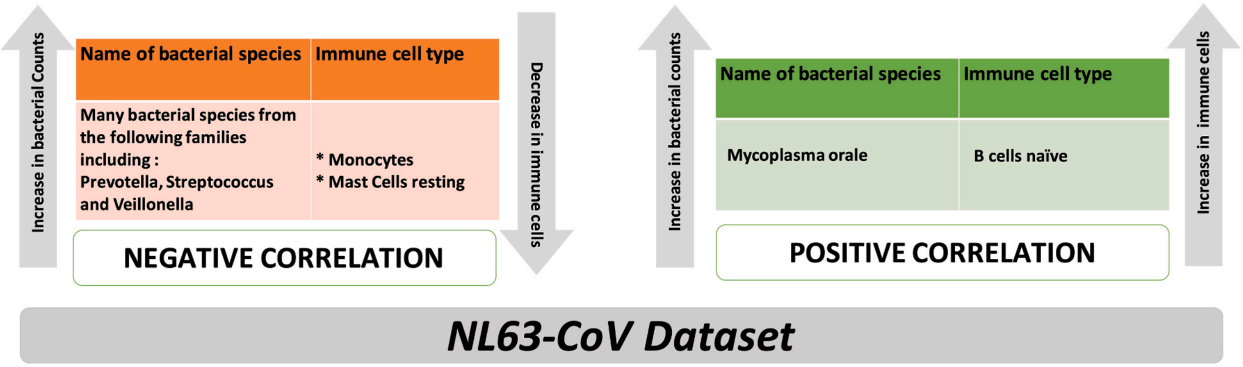
Species and Region name	Immune cell type
NL63 coronavirus	N/A
Human coronavirus 229E – Gene S; product=surface glycoprotein; Name=NP_073551.1	Monocytes
Human coronavirus 229E – Name=NP_073549.1; product=replicase polyprotein 1ab	Monocytes
Human coronavirus 229E – Name=NP_073550.1; product=replicase polyprotein 1a	Monocytes
Human coronavirus 229E – Gene HCoV229Egp1	Monocytes

Gene level Analysis
NEGATIVE CORRELATION

Species and Region name	Immune cell type
NL63 coronavirus	N/A
Human coronavirus 229E – Name=NP_073549.1; product=replicase polyprotein 1ab	B cells naïve
Human coronavirus 229E – Name=NP_073549.1; product=replicase polyprotein 1ab	NK cells activated
Human coronavirus HKU1 – Gene orf1ab	Mast cells resting
Human coronavirus HKU1 – Name=YP_173236.1; gene=orf1ab; product=orf1ab polyprotein	Mast cells resting

Gene Level Analysis
POSITIVE CORRELATION

Fig. 6. Summary of Statistically significant correlation between viral gene expression and immunological cell types for 4 NL63-CoV patients. (A) Viral genome level correlation between viral load and fraction of immune cells (B). Viral gene level correlation between viral gene expression vs fraction of immune cells.



Name of bacterial species	Immune cell type
Many bacterial species from the following families including : Prevotella, Streptococcus and Veillonella	* Monocytes * Mast Cells resting

NEGATIVE CORRELATION

Name of bacterial species	Immune cell type
Mycoplasma orale	B cells naïve

POSITIVE CORRELATION

Fig. 7. Summary of correlation analysis between bacterial abundance with immunological cell types for the NL63-CoV dataset.

including *Prevotella, Streptococcus and Veillonella* (Fig. 7). The full correlation matrix results for this analysis are provided as Supplementary File 5B

3.2.6. *Correlation between viral load (genomic level) with metagenomic abundance*

We chose a p-value cut off of 0.05 to get a list of 15 statistically significant features that correlated viral load (genomic level) with metagenomic abundance. We focused on the results in the NL63 coronavirus genome. We found no bacterial species positively correlated with the NL63 coronavirus genome. There were a few bacterial species negatively correlated with the NL63 coronavirus genome including *Fusobacterium, Prevotella, Streptococcus, Treponema and Veillonella (Fig. 7)*. The full correlation matrix results for this analysis are provided as Supplementary File 5D.

4. Discussion

Our goal was to explore the multidimensional landscape of infected lung tissue microenvironment to better understand complex interactions between virus, immune response and microbiome in the lungs of COVID-19 patients in comparison with NL63-CoV patients. By utilizing three types of machine learning based bioinformatics tools, we were able to detect and quantitate three different fractions of short reads from RNA-seq data files: fraction of viral RNA (at genome and gene level), Human RNA (transcripts and gene level) and bacterial RNA (metagenomic analysis).

4.1. *Immune landscape and correlation analysis in the SARS-CoV-2 dataset*

Our correlation analysis of the three types of measurements in the *SARS-CoV-2 dataset* has shown significant correlation between viral load as well as the level of specific viral gene expression with the fractions of immune cells present in lung lavage as well as with the abundance of major fractions of the lung microbiome.

We saw from our analysis of the SARS-CoV-2 dataset that Macrophages M2 are positively correlated with viral load (Fig. 4A). Macrophages are special immune cells that detect, ingest and destroy target cells. It works by stimulating the immune system through M1 macrophages, and can also encourage tissue repair with the help of M1 macrophages [43,44]. The macrophages located in the lungs include *alveolar and interstitial macrophages*. If these macrophages are excessively activated, they could create a ‘cytokine storm’ which leads to the release of pro-inflammatory factors such as interleukins, and hyper-inflammation; and causes immune cells to attack the organs in the body [45,46]. Such a cytokine storm was commonly found in severe COVID-19 patients with Acute respiratory

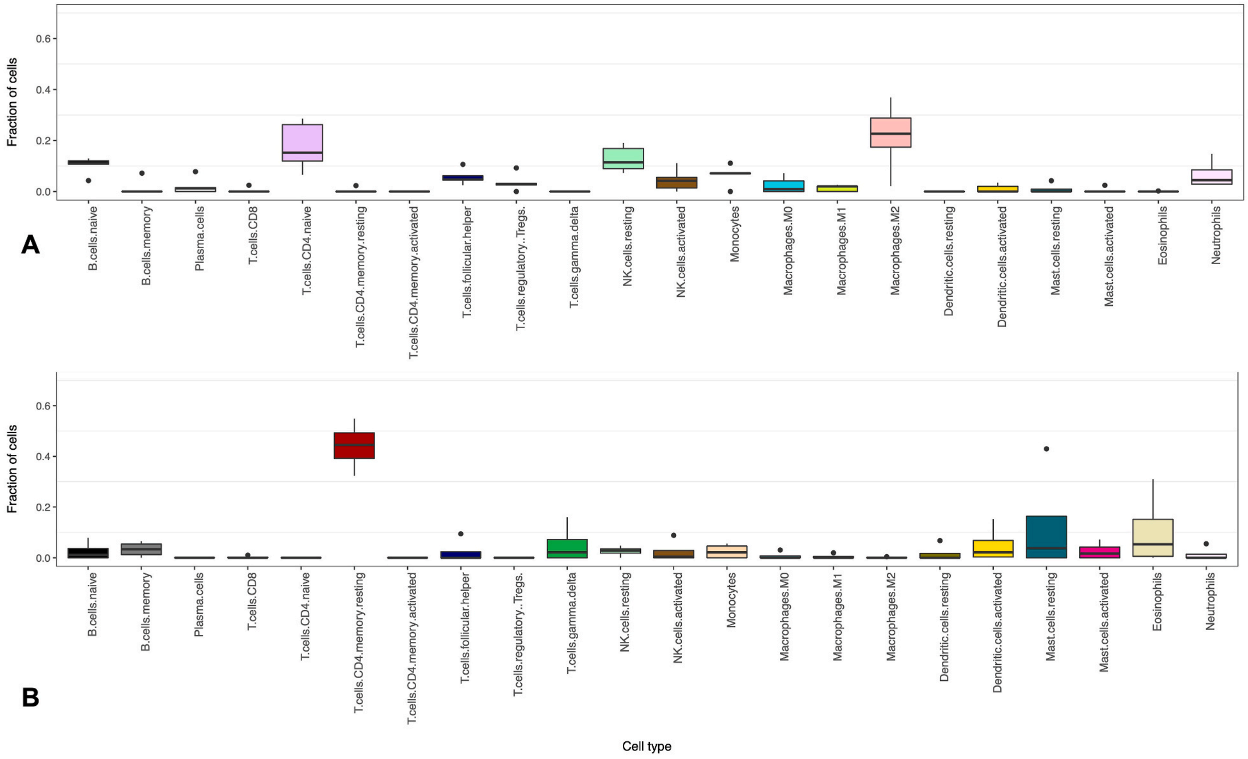


Fig. 8. Box plot for 22 types of immune cells in (a) 5 patients from the SARS-COV-2 dataset and (b) 4 patients from the NL63-COV-Hiseq dataset (Sample 1 from this cohort was ignored for this box plot due to low coverage).

distress syndrome (ARDS) [46]. Recent publications have found that Macrophages M1 in the lung amplify and spread the COVID-19 viral infection, while Macrophages M2 degrade the SARS-CoV-2 virus and limit the spread of infection [47]. Scientists have been exploring treatments that target the pathways in order to lessen the cytokine storm effect in severe COVID-19 disease and some clinical trials are now underway [48].

Another interesting result was the inverse correlation of NK cells activated with viral load (Fig. 4 A). In other words, the NK cells were inactivated in our analysis of the *SARS-CoV-2 dataset*. Natural killer (NK) cells are known to be one of the most important members of the innate immune response and fight against infected cells [49]. Impairment of cells and its inability to fight or kill the infected cells is well known in COVID [50]. Jewett et al. [51] also found a correlation between reduction of NK cells and severity of disease [51]. Such a depletion pattern of NK cells was also observed in patients affected with SARS virus [52]. NK cells are one of the most abundant types of lymphocytes in the lung, and could offer an interesting venue for COVID-19 treatments. Scientists are currently exploring new treatments that use genetically engineered CAR-NK cells that may not trigger a cytokine storm [50]. Clinical trials are ongoing (NCT04324996 and NCT04634370) and scientists are now exploring immunotherapy and enhancement-based treatments using NK cells to combat the disease progression and severity of COVID-19 [51,53–55].

Berentschot et al. studied fatigue in Long COVID patients and found severity in fatigue to be associated with stronger monocyte activation confirming that the study of immune dysregulation is critical in Long COVID patients [56].

Bacterial co-infections were not very frequent, but more commonly found in critically ill COVID-19 patients [57–59]. Bacterial pathogens found in the lungs of COVID-19 patients included *Enterobacter species, Pseudomonas,* Streptococcus pneumoniae were also found in our results (Fig. 6). The *Enterobacteriaceae* species was found to be resistant to antibiotics in some COVID-19 patients [58]. Although not common, such infections in COVID-19 patients were complex to treat since it was not easy to distinguish bacterial co-infections from viral infections of the respiratory tract. Our correlation results also showed a high correlation of these bacteria with immune markers including Eosinophils and activated NK cells (Fig. 5A). This matches findings from Mason et al. [60] who recommended studying inflammatory markers including Lymphocytes (such as NK cells, T cells, and B cells), and Neutrophils to distinguish the bacterial co-infections from viral infections of the respiratory tract [60]. Our results also showed a positive correlation of SARS-CoV-2 viral load with *Schaalia odontolytica* bacteria (Fig. 5B), and similar results were seen in Zhou et al. [61].

4.2. Immune landscape and correlation analysis in the NL63-CoV dataset

Even though the SARS-CoV-2 and the NL63-CoV viruses were from different groups, SARS-CoV-2 virus was from group ß and NL63 from group Alpha, they used the same angiotensin-converting enzyme 2 (ACE2) receptor for binding the spike protein for cell entry. In some studies, NL63 has been used as a surrogate to study the SARS-CoV-2 virus [62,63].

We examined the immune landscapes of the patients in the *SARS-CoV-2 and NL63-CoV* datasets with the help of box plots (Fig. 8A and B). It indicated that Macrophages M2, T cells CD4 naïve, and Natural Killer (NK) cells were the most abundant in the immune landscape of the *SARS-CoV-2* dataset. On the other hand, we saw the dominance of mast cells and CD4 memory resting T cells in our analysis of the NL63-CoV dataset. Richards et al. [64] examined circulating T cells in human endemic coronaviruses (hCoV) including HCoV-229E, NL63-CoV, OC43 and HKU1; and theorized that the *memory CD4 T-cells* found in patients exposed to infection from these endemic strains could also influence the immune response to SARS-CoV-2 infection and vaccination [64,65].

4.3. Findings from analyses on both SARS-CoV-2 and NL63-CoV datasets

In a normal tissue, T cells including CD4 and CD8 work to identify antigens from foreign pathogens. When such an event happens, the T cells differentiate into short lived *effector T cells* that work to control the foreign pathogens. In the long term, the effector T cells are lost, but the memory T-cells are preserved to enable long-term immune response [66]. In published articles, Effector B and T cells were found elevated, and associated with the long-term side effects of the disease [64] [65], [67], [68]. In the analysis of the *NL63-CoV dataset*, we saw an elevated fraction of the *memory CD4 T-cells* that could potentially help with the long-term immune response to the future SARS-CoV-2 infection [64,65]. This finding indicates that a treatment or a vaccine that could potentially mediate the T-cell response to produce effective *memory CD4 T-cells;* and control the levels of *effector T-cell activity* may not only provide immunity from the SARS-CoV-2 infection, but also help prevent the adverse side effects in Long COVID in patients affected by this virus [66,69].

Shaath et al. [70] examined the blood samples from COVID-19 patients admitted to the hospital intensive care unit (ICU) vs those not in ICU and found several mRNA based markers associated with severity of the COVID-19 disease. The authors found several pathways related to NK cells and interferon signaling to be downregulated in ICU COVID-19 patients. The authors recommended restoration of NK cells, and mediation of interferon-gamma as potential therapeutic options [70]. This work was done on blood samples and hence their findings are valid for a systemic immune response. These findings at the systemic level are in agreement with our findings indicating a similar type of immune response at the local tissue level in the lungs.

4.4. Relevance of machine learning tools and algorithms

As COVID-19 cases continue to rise around the world, researchers are harnessing the computational power of machine learning and artificial intelligence (AI) tools to not only create prediction and diagnostic tools for COVID-19 [71,72] but also improve outcomes [73]. In this paper, we described the application of machine learning tools to process the raw sequencing data generated by NGS technology, and also explore the immune, viral and bacterial landscape of the *SARS-CoV-2 and NL63-CoV datasets*.

While traditional laboratory techniques allow direct detection of immune cells in the blood, it is more difficult to do for other types of tissues where immune cells can be detected using a technique called flow cytometry or image cytometry methods. But both of these techniques are difficult and labor- and time-consuming.

RNA-sequencing (RNA-seq) technology in conjunction with the application of machine learning based virtual flow cytometry tools could be considered a potential alternative. Such an in-silico process would enable researchers to not only estimate the immune cell environment, but also work towards new hypotheses and therapies that would mediate appropriate immune response using T cells for long term immunity; and help to minimize adverse side effects from the SARS-CoV-2 infection [74,75].

4.5. Relevance of the multi-modal RNA-seq based analysis of samples from COVID patients

In recent news, researchers believed that reducing the intensity of the immune system response could alleviate symptoms in some Long COVID patients [76]. Research has also implicated the microbiome affecting inflammation [77], and changed humoral responses and antibody responses [78]. It indicates that the interplay of the viral infection, immune system and microbiome has to be studied in detail to enable new and improved treatments of long-COVID. The multimodal RNA-seq based analysis presented here provides unique capabilities of simultaneous analysis of these interactions in a single patient sample.

Our exploratory study has also uncovered novel insights into complex regulatory signaling interactions and correlative patterns between the viral infection, inhibition of innate and adaptive immune response as well as microbiome landscape of the lung tissue. Many of our findings from the analysis of the immune landscape of these two datasets, along with correlation analysis have been corroborated in published literature.

4.6. Challenges and limitations

At the beginning of this multi-modal proof of concept project, we were unsure of what the data would reveal; but at the end of this phase, we believe our multi-model RNA-seq based approach is worth pursuing. While we recognize that our sample size is small compared to the vast number of affected individuals, it aligns with the nature of a proof-of-concept research. Moreover, challenges faced by the data generators of the *NL63 dataset* regarding the partial genome sequences obtained during sequencing by NGS methods, were reflected in the data analysis, some of which did not yield significant results.

4.7. Future directions

Our proof-of-concept highlights the potential for various analyses achievable with bulk RNA-seq data. Future directions include expanding the scope of our study to include a larger cohort of COVID-19 patients. This will increase the statistical power and facilitate the incorporation of additional molecular data types, and enabling a more comprehensive multi-modal analysis exploring diverse microenvironments.

In published articles, elevated levels of Effector B and T cells have been consistently associated with the long-term side effects of the disease [64] [65], [67], [68]. These studies confirm that observing and mediating immune response, including T cell responses could be critical in the treatment of Long COVID. Building upon these insights, we plan to expand our multi-modal analysis to a deeper immune profile in the future.

Scientists are also researching the role of autoantibodies and their possible link to the long-term side effects of Long COVID [79]. Autoantibodies are antibodies i.e. immune proteins that target a person's own cells and organs, and cause autoimmune diseases such as Lupus [80]. Following along those lines, we plan to explore the immune proteomic landscape in the future to improve our understanding of Long COVID.

Furthermore, while some COVID-19 patients exhibited a return to normal immune profiles post-recovery, others displayed persistent alterations in immune cell populations months after clearing the SARS-CoV-2 virus [67], [81–83]. Comparing the immune profiles of these two groups of patients could also help shed light on the complexities of this disease.

5. Conclusion

In this methodology paper, we applied multiple machine learning tools to NGS data analysis of lung tissue samples from COVID-19 patients. We explored the SARS-CoV-2 gene expression patterns and compared it with another endemic coronavirus, NL63. Finally, we explored the immunological landscape of the lung microenvironment from the *SARS-Cov-2 and* nasopharyngeal microenvironment from the *NL63-CoV datasets.*

Our exploratory study has provided novel insights into complex regulatory signaling interactions and correlative patterns between the viral infection, inhibition of innate and adaptive immune response as well as microbiome landscape of the lung tissue. Many of our findings from the analysis of the immune landscape of these two datasets, along with correlation analysis have been corroborated in published literature proving that the study of the immune system warrants further analysis and exploration.

The study of how the SARS-CoV-2 virus interacts with the immune system; and comparing and contrasting the immune system in patients affected by endemic viruses could offer important insights into immunoprotecting against SARS-CoV-2; and shed light on new therapies to combat severe COVID-19 disease.

This proof-of-concept case study demonstrates the possibilities of the various types of analyses that can be performed from this type of RNA-seq data. These initial findings on a small group of samples could provide a better understanding of the diverse dynamics of immune response and the side effects of the SARS-CoV-2 infection but require further validation on a larger cohort of samples.

Data availability statement

The datasets used in this manuscript are available online.

***SARS-CoV-2* Dataset:** We downloaded RNA-seq data from 5 patients affected with the SARS COV 2 virus from the NCBI SRA data repository PRJNA605983 (SRP249613) [24,25]. These 5 patients were from the early stage of the Wuhan seafood market pneumonia virus outbreak in China. The downloaded data were raw sequences in the form of.FASTQ files. Total RNA was extracted from bronchoalveolar lavage fluid and then next generation sequencing (NGS) was performed using the Illumina platform. Out of the 5 samples, 4 were profiled on the Illumina HiSeq 3000 platform, and one on the Illumina HiSeq 1000 platform.

Human NL63 coronavirus dataset (referred to as the *NL63-CoV* dataset in this paper):

We found a public dataset was from 5 pediatric patients with severe lower respiratory infection by NL63 coronavirus with deep sequencing data performed on Illumina HiSeq platform. The downloaded data were raw sequences in the form of.FASTQ files obtained from NCBI PRJA601331 [26,27].

Funding

The work was partly funded and in part supported by the National Center For Advancing Translational Sciences (NCATS) of the National Institutes of Health (NIH) under Award Number UL1TR001409.

CRediT authorship contribution statement

Krithika Bhuvaneshwar: Writing – original draft, Validation, Methodology, Formal analysis, Conceptualization. **Subha Madhavan:** Writing – review & editing. **Yuriy Gusev:** Writing – review & editing, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to acknowledge the CGC Seven Bridges team for enabling some of the high-throughput analysis using the CENTRIFUGE metagenomic tool.

List of abbreviations

COVID-19	The coronavirus disease 2019
SARS-CoV-2	SARS coronavirus 2
SARS-CoV	SARS coronavirus
IFN	Type I interferons
NCATS	NIH National Center for Advancing Translational Sciences
EHRs	Rlectronic health records
WSI	Whole slide images
PASC	Post-acute sequelae of SARS-CoV-2 infection
HCoV-NL63	NL63 coronavirus
hCoV	Human endemic coronaviruses
NGS	Next generation sequencing
SIRS	Severe acute respiratory infection
SIMD	Single-instruction multiple-data
SB	Seven Bridges
CGC	Cancer Genomics Cloud (CGC) Platform
ν-SVR	Nu-linear support vector regression
SVM	Support vector machine
BWT	Burrows-Wheeler transform
FM	Ferragina-Manzini (FM) index
NK	Natural Killer (NK) cells
AI	Srtificial intelligence (AI)
RNA-seq	RNA-sequencing
ML	Machine learning
AI	Artificial Intelligence

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e32772.

References

[1] COVID-19 Coronavirus Pandemic, Secondary COVID-19 coronavirus pandemic. https://www.worldometers.info/coronavirus/, 2021. (Accessed 30 March 2021).
[2] I. Al-Jahdhami, K. Al-Naamani, A. Al-Mawali, The post-acute COVID-19 syndrome (long COVID), Oman Med. J. 36 (1) (2021) e220.
[3] COVID-19 rapid resource center > age distribution and critical illness severity of COVID-19 secondary COVID-19 rapid resource center > age distribution and critical illness severity of COVID-19 2021. https://www.sccm.org/COVID19RapidResources/Resources/Age-Distribution-and-Critical-Illness-Severity-of. (Accessed 30 March 2021).
[4] Q. Ma, J. Liu, Q. Liu, et al., Global percentage of asymptomatic SARS-CoV-2 infections among the tested population and individuals with confirmed COVID-19 diagnosis: a systematic review and meta-analysis, JAMA Netw. Open 4 (12) (2021) e2137257.
[5] WHO. Global research on coronavirus disease (COVID-19). Secondary Global research on coronavirus disease (COVID-19). https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov. Last Accessed April 24 2024.
[6] N. Communications, Top 25 COVID-19 articles of 2022. Secondary top 25 COVID-19 articles of 2022. https://www.nature.com/collections/ahdejdadbd, 2023. (Accessed 24 April 2024).
[7] N. Communications, The top 25 COVID-19 articles of 2023. Secondary the top 25 COVID-19 articles of 2023. https://www.nature.com/collections/gjajjhfcha, 2024. (Accessed 24 April 2024).
[8] Nature S. Coronavirus Collections.
[9] PMC. PMC COVID-19 Collection, Secondary PMC COVID-19 collection. https://www.ncbi.nlm.nih.gov/pmc/about/covid-19/. (Accessed 25 April 2024).
[10] V.V. Khanna, K. Chadaga, N. Sampathila, et al., A machine learning and explainable artificial intelligence triage-prediction system for COVID-19, Decision Analytics Journal 7 (4) (2023).
[11] K. Chadaga, S. Prabhu, V. Bhat, et al., An explainable multi-class decision support framework to predict COVID-19 prognosis utilizing biomarkers, Cogent Engineering 10 (2) (2023).
[12] K. Chadaga, S. Prabhu, V. Bhat, et al., Artificial intelligence for diagnosis of mild-moderate COVID-19 using haematological markers, Ann. Med. 55 (1) (2023) 2233541.
[13] A. Nalbandian, K. Sehgal, A. Gupta, et al., Post-acute COVID-19 syndrome, Nat Med 27 (2021) 601–615.
[14] O. Moreno-Perez, E. Merino, J.M. Leon-Ramirez, et al., Post-acute COVID-19 syndrome. Incidence and risk factors: a Mediterranean cohort study, J. Infect. 82 (3) (2021) 378–383.
[15] S.A. Candan, N. Elibol, A. Abdullahi, Consideration of prevention and management of long-term consequences of post-acute respiratory distress syndrome in patients with COVID-19, Physiother. Theory Pract. 36 (6) (2020) 663–668.
[16] Predicting 'long COVID syndrome' with help of a smartphone app. Secondary predicting 'long COVID syndrome' with help of a smartphone app. https://directorsblog.nih.gov/tag/post-acute-sequelae-of-covid-19/. (Accessed 30 March 2021).
[17] US health agency will invest $1 billion to investigate 'long COVID'. Secondary US health agency will invest $1 billion to investigate 'long COVID' 2021. https://www.nature.com/articles/d41586-021-00586-y. (Accessed 30 March 2021).
[18] £18.5 million awarded to new research projects to understand and treat long COVID. Secondary £18.5 million awarded to new research projects to understand and treat long COVID 2021. https://www.nihr.ac.uk/news/185-million-awarded-to-new-research-projects-to-understand-and-treat-long-covid/26895. Last Accessed March 30, 2021.
[19] H.E. Davis, L. McCorkell, J.M. Vogel, et al., Author Correction: long COVID: major findings, mechanisms and recommendations, Nat. Rev. Microbiol. 21 (6) (2023) 408.
[20] A.J. Westermann, J. Vogel, Cross-species RNA-seq for deciphering host-microbe interactions, Nat. Rev. Genet. 22 (6) (2021) 361–378.
[21] A.J. Westermann, S.A. Gorski, J. Vogel, Dual RNA-seq of pathogen and host, Nat. Rev. Microbiol. 10 (9) (2012) 618–630.
[22] S. Abdul-Rasool, B.C. Fielding, Understanding human coronavirus HCoV-NL63, Open Virol. J. 4 (2010) 76–84.
[23] K.A. Richards, M. Glover, J.C. Crawford, et al., Circulating CD4 T cells elicited by endemic coronaviruses display vast disparities in abundance and functional potential linked to antigen specificity and age, J. Infect. Dis. 2 (2021).
[24] NCBI SRA SRP249613, Secondary NCBI SRA SRP249613. https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP249613. (Accessed 31 March 2021).
[25] P. Zhou, X.L. Yang, X.G. Wang, et al., A pneumonia outbreak associated with a new coronavirus of probable bat origin, Nature 579 (7798) (2020) 270–273.
[26] Respiratory Sample from Hospitalized Children with SARI (PRJA601331).
[27] L. Zhang, M. Gan, Z. Zhang, et al., Complete genome sequences of five human coronavirus NL63 strains causing respiratory illness in hospitalized children in China, Microbiol Resour Announc 9 (8) (2020).
[28] NCBI Virus. Secondary NCBI Virus. https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&CollectionDate_dr=1950-01-01T00:00:00Z%20TO%20NOW&CreateDate_dt=1950-01-01T00:00:00Z%20TO%20NOW&HostLineagess=Homo%20sapiens%20(human),%20taxid:9606&SourceDBs=RefSeq&Completenesss=complete. Last Accessed April 5, 2021.
[29] K. Bhuvaneshwar, L. Song, S. Madhavan, et al., viGEN: an open source pipeline for the detection and quantification of viral RNA in human tumors, Front. Microbiol. 9 (2018) 1172.
[30] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with Bowtie 2, Nat. Methods 9 (4) (2012) 357–359.
[31] D. Kim, L. Song, F.P. Breitwieser, et al., Centrifuge: rapid and sensitive classification of metagenomic sequences, Genome Res. 26 (12) (2016) 1721–1729.
[32] B. Li, C.N. Dewey, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome, BMC Bioinf. 12 (2011) 323.
[33] Correction: the cancer genomics Cloud: collaborative, reproducible, and democratized-A new paradigm in large-scale computational research, Cancer Res. 78 (17) (2018) 5179.
[34] J.W. Lau, E. Lehnert, A. Sethi, et al., The cancer genomics Cloud: collaborative, reproducible, and democratized-A new paradigm in large-scale computational research, Cancer Res. 77 (21) (2017) e3–e6.
[35] B. Chen, M.S. Khodadoust, C.L. Liu, et al., Profiling tumor infiltrating immune cells with CIBERSORT, Methods Mol. Biol. 1711 (2018) 243–259.
[36] A.M. Newman, C.L. Liu, M.R. Green, et al., Robust enumeration of cell subsets from tissue expression profiles, Nat. Methods 12 (5) (2015) 453–457.
[37] Centrifuge Classifier for metagenomic sequences, Secondary Centrifuge classifier for metagenomic sequences. https://ccb.jhu.edu/software/centrifuge/. (Accessed 7 April 2019).
[38] F.P. Breitwieser, S.L. Salzberg, Pavian: interactive analysis of metagenomics data for microbiome studies and pathogen identification, Bioinformatics 36 (4) (2020) 1303–1304.
[39] Pavian, Secondary pavian. https://github.com/fbreitwieser/pavian. (Accessed 7 April 2021).
[40] R Core Team, R: A Language and Environment for Statistical Computing, R: A language and environment for statistical computing, 2017. Secondary R Core Team (2017), https://www.R-project.org/. (Accessed 7 April 2021).
[41] Severe acute respiratory syndrome (SARS), Secondary Severe acute respiratory syndrome (SARS). https://www.cdc.gov/sars/index.html. (Accessed 9 April 2021).

[42] V. Subramanyam, D. Paramshivan, A. Kumar, et al., Using Sankey diagrams to map energy flow from primary fuel to end use, Energy Convers. Manag. 91 (2015) 342–352.
[43] Macrophages, Secondary macrophages. https://www.immunology.org/public-information/bitesized-immunology/cells/macrophages. (Accessed 19 April 2021).
[44] Wikipedia. Macrophages. Secondary Macrophages. https://en.wikipedia.org/wiki/Macrophage. Last Accessed April 19, 2021.
[45] S. Meidaninikjeh, N. Sabouni, H.Z. Marzouni, et al., Monocytes and macrophages in COVID-19: friends and foes, Life Sci. 269 (2021) 119010.
[46] P. Song, W. Li, J. Xie, et al., Cytokine storm induced by SARS-CoV-2, Clin. Chim. Acta 509 (2020) 280–287.
[47] J. Lv, Z. Wang, Y. Qu, et al., Distinct uptake, amplification, and release of SARS-CoV-2 by M1 and M2 alveolar macrophages, Cell Discov 7 (1) (2021) 24.
[48] J. Zhang, H. Wu, X. Yao, et al., Pyroptotic macrophages stimulate the SARS-CoV-2-associated cytokine storm, Cell. Mol. Immunol. 5 (2021) 1305–1307.
[49] Wikipedia. Natural Killer Cell, Secondary natural killer cell. https://en.wikipedia.org/wiki/Natural_killer_cell. (Accessed 20 April 2021).
[50] A. Zafarani, M.H. Razizadeh, S. Pashangzadeh, et al., Natural killer cells in COVID-19: from infection, to vaccination and therapy, Future Virol (2023), https://doi.org/10.2217/fvl-2022-0040.
[51] A. Jewett, The potential effect of novel coronavirus SARS-CoV-2 on NK cells; A perspective on potential therapeutic interventions, Front. Immunol. 11 (2020) 1692.
[52] National Research Project for SARS BG, The involvement of natural killer cells in the pathogenesis of severe acute respiratory syndrome, Am. J. Clin. Pathol. 121 (4) (2004) 507–511.
[53] C. Bao, X. Tao, W. Cui, et al., Natural killer cells associated with SARS-CoV-2 viral RNA shedding, antibody response and mortality in COVID-19 patients, Exp. Hematol. Oncol. 10 (1) (2021) 5.
[54] ClinicalTrials.gov, NCT04634370: fase I clinical trial on NK cells for COVID-19. Secondary NCT04634370: fase I clinical trial on NK cells for COVID-19. https://clinicaltrials.gov/ct2/show/NCT04634370. (Accessed 20 April 2021).
[55] Boosting natural killer cells for the treatment of COVID-19. Secondary boosting natural killer cells for the treatment of COVID-19 2021. https://pediatricsnationwide.org/. (Accessed 20 April 2021). https://pediatricsnationwide.org/2021/01/05/boosting-natural-killer-cells-for-the-treatment-of-covid-19/.
[56] J.C. Berentschot, H.A. Drexhage, D.G. Aynekulu Mersha, et al., Immunological profiling in long COVID: overall low grade inflammation and T-lymphocyte senescence and increased monocyte activation correlating with increasing fatigue severity, Front. Immunol. 14 (2023) 1254899.
[57] R.F. Alam, M.M. Hossain, A.B. Ibne Momen, et al., COVID-19 with multiple bacterial Co-infections: a case report, European Journal of Medical and Health Sciences 3 (1) (2021) 1–4.
[58] H. Mahmoudi, Bacterial co-infections and antibiotic resistance in patients with COVID-19, GMS Hyg Infect Control 15 (2020) Doc35.
[59] B.J. Langford, M. So, S. Raybardhan, et al., Bacterial co-infection and secondary infection in patients with COVID-19: a living rapid review and meta-analysis, Clin. Microbiol. Infect. 26 (12) (2020) 1622–1629.
[60] C.Y. Mason, T. Kanitkar, C.J. Richardson, et al., Exclusion of bacterial co-infection in COVID-19 using baseline inflammatory markers and their response to antibiotics, J. Antimicrob. Chemother. (5) (2021) 1323–1331.
[61] T. Zhou, J. Wu, Y. Zeng, et al., SARS-CoV-2 triggered oxidative stress and abnormal energy metabolism in gut microbiota, MedComm 3 (1) (2020) e112, 2022.
[62] A. Chakraborty, A. Diwan, NL63: a better surrogate virus for studying SARS-CoV-2, Integrative Molecular Medicine 8 (2) (2020).
[63] G. Castillo, R.K. Nelli, K.S. Phadke, et al., SARS-CoV-2 is more efficient than HCoV-NL63 in infecting a small subpopulation of ACE2+ human respiratory epithelial cells, Viruses 15 (3) (2023).
[64] K.A. Richards, M. Glover, J.C. Crawford, et al., Circulating CD4 T cells elicited by endemic coronaviruses display vast disparities in abundance and functional potential linked to both antigen specificity and age, J. Infect. Dis. (9) (2021) 1555–1563.
[65] H.X. Tan, W.S. Lee, K.M. Wragg, et al., Adaptive immunity to human coronaviruses is widespread but low in magnitude, Clin Transl Immunology 10 (3) (2021) e1264.
[66] N.N. Jarjour, D. Masopust, S.C. Jameson, T cell memory: understanding COVID-19, Immunity 54 (1) (2021) 14–18.
[67] T. Bilich, A. Nelde, J.S. Heitmann, et al., T cell and antibody kinetics delineate SARS-CoV-2 peptides mediating long-term immune responses in COVID-19 convalescent individuals, Sci. Transl. Med. 590 (2021).
[68] UKRI, The immune system and long COVID. Secondary the immune system and long COVID March 1, 2021 2021. https://www.ukri.org/our-work/tackling-the-impact-of-covid-19/understanding-coronavirus-covid-19-and-epidemics/the-immune-system-and-long-covid/. (Accessed 21 April 2021).
[69] T cells found in COVID-19 patients 'bode well' for long-term immunity. Secondary T cells found in COVID-19 patients 'bode well' for long-term immunity, May. 14, 2020, https://www.sciencemag.org/news/2020/05/t-cells-found-covid-19-patients-bode-well-long-term-immunity, 2020. (Accessed 22 April 2021).
[70] H.S.N.M. Alajez, Identification of PBMC-based molecular signature associational with COVID-19 disease severity, Heliyon 5 (2021) e06866.
[71] FDA Approves Machine Learning Tool for COVID-19 Screening, Secondary FDA approves machine learning tool for COVID-19 screening march 23, 2021. https://healthitanalytics.com/news/fda-approves-machine-learning-tool-for-covid-19-screening, 2021. (Accessed 22 April 2021).
[72] Y. Zoabi, S. Deri-Rozov, N. Shomron, Machine learning-based prediction of COVID-19 diagnosis based on symptoms, NPJ Digit Med 4 (1) (2021) 3.
[73] AI, Machine learning tools help predict COVID-19 outcomes. Secondary AI, machine learning tools help predict COVID-19 outcomes dec 21. https://healthitanalytics.com/news/ai-machine-learning-tools-help-predict-covid-19-outcomes, 2020 2020. (Accessed 22 April 2021).
[74] Y.A. Lyons, S.Y. Wu, W.W. Overwijk, et al., Immune cell profiling in cancer: molecular approaches to cell-specific identification, npj Precis. Oncol. 1 (1) (2017) 26.
[75] Testing a new COVID-19 test: how T-cells beat antibodies in helping to detect past infections, Secondary Testing a new COVID-19 test: How T-cells beat antibodies in helping to detect past infections February 26 (2021). https://www.geekwire.com/2021/testing-new-covid-19-test-t-cells-beat-antibodies-detecting-past-infections/. (Accessed 22 April 2021).
[76] NPR, New clues to the biology of long COVID are starting to emerge. Secondary New clues to the biology of long COVID are starting to emerge 2021. https://www.npr.org/sections/health-shots/2021/11/12/1053509795/long-covid-causes-treatment-clues. (Accessed 8 December 2021).
[77] J.P. Haran, E. Bradley, A.L. Zeamer, et al., Inflammation-type dysbiosis of the oral microbiome associates with the duration of COVID-19 symptoms and long COVID, JCI Insight 6 (20) (2021).
[78] J. Klein, J. Wood, J. Jaycox, et al., Distinguishing features of Long COVID identified through immune profiling, Nature 623 (2023) 139–148.
[79] R. Khamsi, Rogue antibodies could be driving severe COVID-19, Nature 590 (7844) (2021) 29–31.
[80] Wikipedia Autoantibody, Secondary autoantibody. https://en.wikipedia.org/wiki/Autoantibody. (Accessed 21 April 2021).
[81] Liu J., Yang X., Wang H., et al., The analysis of the long-term impact of SARS-CoV-2 on the cellular immune system in individuals recovering from COVID-19 reveals a profound NKT cell impairment, mBio. 2021 Apr 27;12(2):e00085-21. doi: 10.1128/mBio.00085-21.
[82] L. Bergamaschi, F. Mescia, L. Turner, et al., Delayed bystander CD8 T cell activation, early immune pathology and persistent dysregulation characterise severe COVID-19, medRxiv (2021).
[83] M.J. Peluso, A.N. Deitchman, L. Torres, et al.. Long-term SARS-CoV-2-specific immune and inflammatory responses across a clinically diverse cohort of individuals recovering from COVID-19, 2021 medRxiv.