

An Integrative Analyzer for Publicly Available COVID-19 Genome-wide Association Studies

Zhongshan Cheng, Center for Applied Bioinformatics, St. Jude Children's Research Hospital, Memphis, TN, 38015, US

ABSTRACT

COVID-19 has led to over 7 million deaths globally, with human genetic factors playing a significant role in the variability of disease severity. To identify genetic contributors to severe COVID-19, we developed the "COVID19_GWAS_Analyzer" using SAS OnDemand for Academics. This genome-wide association study (GWAS) analyzer facilitates comprehensive explorative analysis of large-scale, publicly available COVID-19 datasets, including GWAS summary statistics, expression quantitative trait loci (eQTL) data, and single-cell expression data. The package incorporates over 400 SAS macros designed for mining these datasets and other related large-scale COVID-19 resources. The primary objective is to support biologists and geneticists in identifying candidate genes for further investigation based on COVID-19 GWAS summary data. Moreover, the flexibility of combining SAS macros within the SAS OnDemand for Academics platform enables detailed exploration of additional GWAS datasets. The "COVID19_GWAS_Analyzer" is freely available on GitHub at https://github.com/chengzhongshan/COVID19_GWAS_Analyzer.

INTRODUCTION

The COVID-19 pandemic has dramatically impacted global health, with clinical outcomes varying widely across populations. Numerous studies suggest that genetic predispositions play a crucial role in determining the severity of COVID-19, with genome-wide association studies (GWAS) uncovering several loci that may contribute to these variations (COVID-19 Host Genetics Initiative, 2021, 2022, 2023; Pairo-Castineira et al. 2023). One of the key challenges in these studies is integrating multiple datasets from different populations to identify ancestry-specific risk factors for severe disease outcomes. To address this challenge, we developed the COVID19_GWAS_Analyzer, a SAS-based tool designed for integrative analysis of publicly available GWAS datasets. The tool automates critical aspects of the analysis pipeline, including data download, quality control, differential SNP effect size comparison, and advanced statistical methods such as eQTL and single-cell RNA-seq analysis.

STRUCTURE OF COVID19_GWAS_ANALYZER

COVID19_GWAS_Analyzer is a SAS-based tool designed for integrative analysis of publicly available GWAS datasets, which is implemented based on the cloud-based platform SAS OnDemand for Academics. By utilizing SAS OnDemand for Academics, COVID19_GWAS_Analyzer ensures accessibility to users with limited computational resources, making it a scalable and powerful solution for researchers exploring the genetic architecture of COVID-19 severity.

COVID19_GWAS_Analyzer can access to publicly available COVID-19-related GWAS data, such as those provided by the COVID-19 Host Genetics Initiative (HGI) and the UK Biobank (see the workflow in Figure 1). For example, the HGI provides a wealth of GWAS summary statistics, with specific datasets comparing hospitalized COVID-19 patients to non-hospitalized ones (HGI-B1) and to the general population (HGI-B2). These datasets can be used to detect SNPs associated with severe COVID-19 outcomes, focusing on ancestry-specific signals that may explain observed disparities in disease severity (as demonstrated by Cheng et al. 2023 and Zhang et al. 2024).

Figure 1 Workflow of COVID19_GWAS_Analyzer

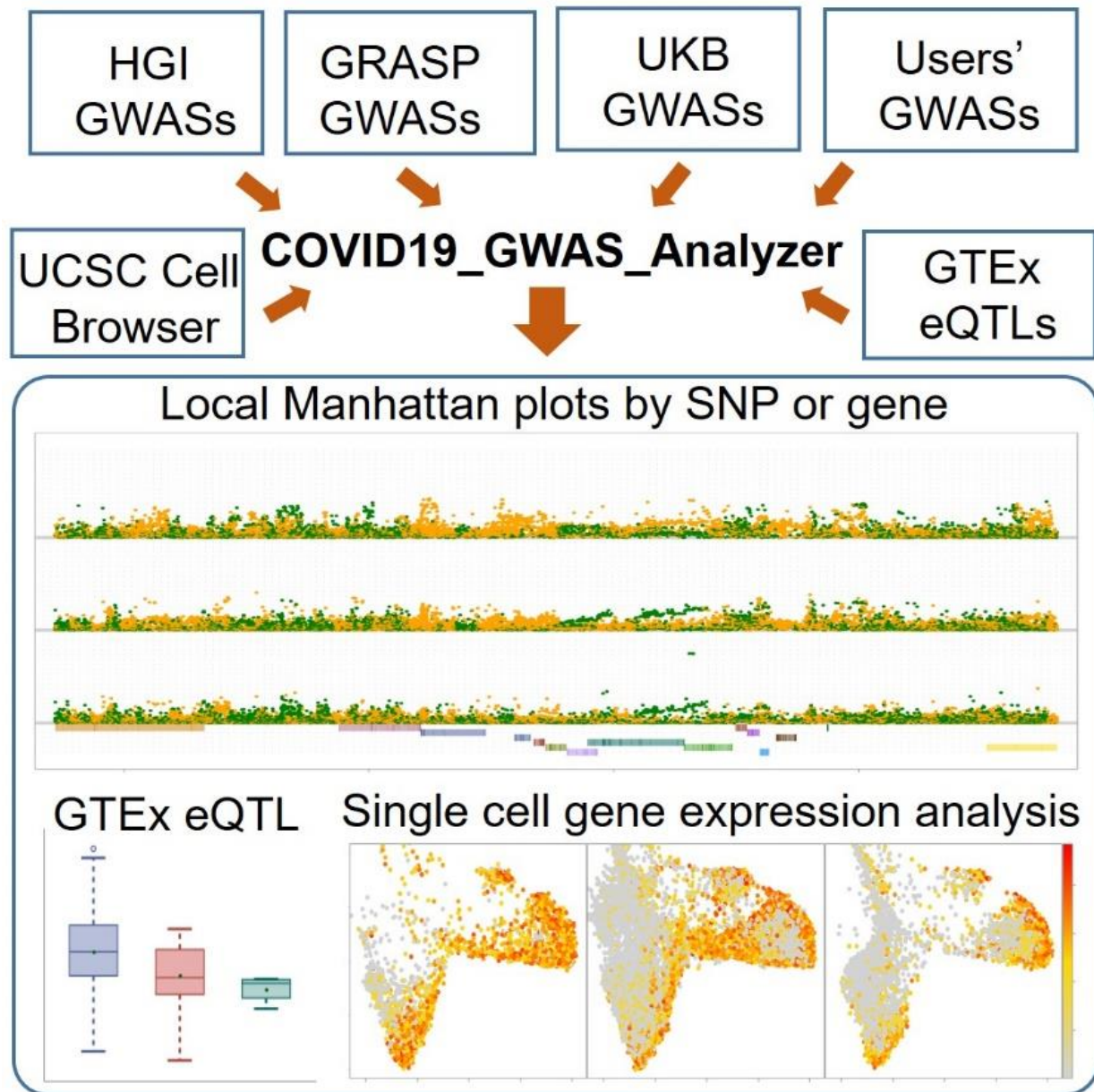


Figure 1. The COVID19_GWAS_Analyzer is a SAS-based protocol designed for integrating and analyzing publicly available genome-wide association studies (GWAS) to investigate genetic factors associated with severe COVID-19. This tool, developed for SAS OnDemand for Academics, automates the process of differential SNP effect size comparisons, expression quantitative trait loci (eQTL) calculations, and single-cell RNA-sequencing analysis.

IMPLEMENTATION OF COVID19_GWAS_ANALYZER

To install and load the necessary SAS macros from COVID19_GWAS_Analyzer, users can execute the following commands in SAS Studio:

```
filename loadmacros url
"https://raw.githubusercontent.com/chengzhongshan/COVID19_GWAS_Analyzer/main/Macros/importallm
acros_ue.sas";
```

```
%include loadmacros;
```

```
%importallmacros_ue(MacroDir=%sysfunc(pathname(HOME))/Macros,fileRgx=.,verbose=0);
```

These macros streamline various aspects of GWAS analysis, including data preparation, statistical comparisons, and visualization. The macro “%importallmacros_ue” ensures that all functions needed for analyses are available.

COMPARISON BETWEEN TWO COVID19 GENOME-WIDE ASSOCIATION STUDIES

The COVID19_GWAS_Analyzer is designed to handle large-scale GWAS data from multiple sources. Here we demonstrate how to compare two COVID-19 hospitalization GWAS datasets (HGI-B1 and HGI-B2). The following macro is used to download and compare the two datasets, focusing on the identification of SNPs that show differential associations with severe COVID-19 outcomes:

```
%GRASP_COVID_Hosp_GWAS_Comparison(
  gwas1=https://grasp.nih.gov/downloads/COVID19GWAS/HGI_B1_ALL.txt.gz,
  gwas2=https://grasp.nih.gov/downloads/COVID19GWAS/HGI_B2_ALL.txt.gz,
  outdir=%sysfunc(pathname(HOME)),
  mk_manhattan_qqplots4twoGWASs=1
);
```

This macro compares the effect sizes of individual SNPs between the two GWAS datasets, identifying variants that show differential associations between hospitalized vs. non-hospitalized individuals (HGI-B1) and hospitalized individuals vs. the general population (HGI-B2). The outputs of this analysis include three SAS datasets saved into the default “HOME” directory in SAS OnDemand for Academics: “GWAS1”, “GWAS2”, and “GWAS1_vs_2”, which can be used to generate Manhattan and Q-Q plots for visualizing the distribution of p-values across the genome.

VISUALIZATION OF GWAS RESULTS

The differential association analysis is visualized using Manhattan plots, which display SNPs across the genome and their corresponding p-values for association with COVID-19 severity. The following SAS macro generates a composite Manhattan plot for the two datasets:

```
%Manhattan4DiffGWASs(
  dsdin=D.GWAS1_vs_2,
  pos_var=pos,
  chr_var=chr,
  P_var=GWAS1_P,
  Other_P_vars=GWAS2_P Pval,
  flip1stGWAS_signal=0
);
```

As shown in Figure 2, the comparison between the HGI-B1 and HGI-B2 datasets reveals several SNPs with suggestively differential associations with COVID-19 severity. Notably, rs16831827, located near the *MAP3K19* gene, shows a large differential effect size between the two datasets, suggesting it may play a role in modulating the severity of COVID-19 in specific populations.

Figure 2 Compact Manhattan plot for 3 GWASs

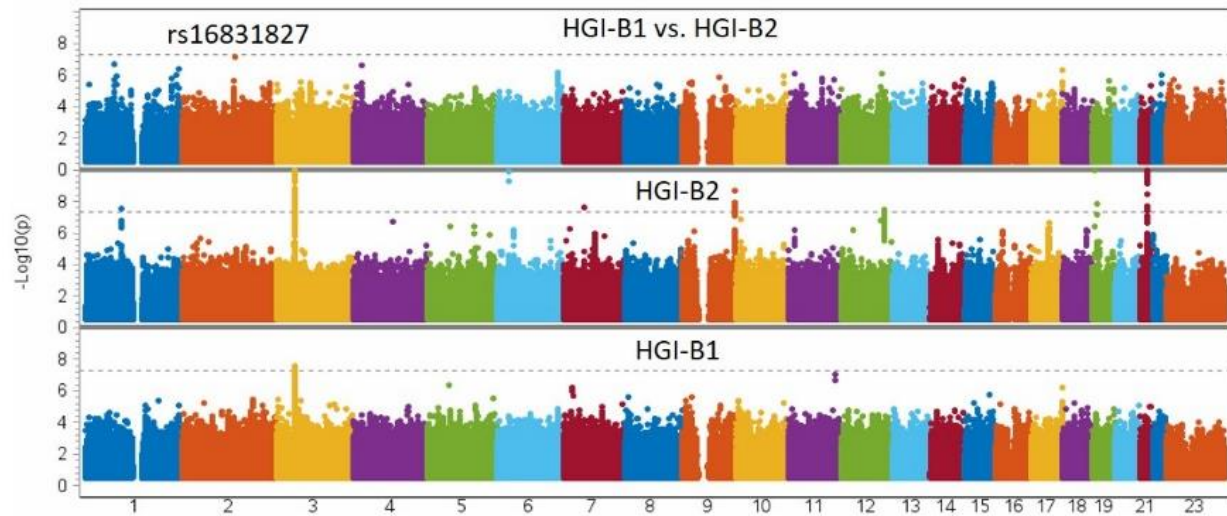


Figure 2. Compact Manhattan plot for 3 GWASs. Three COVID-19 GWASs, including HGI-B1, HGI-B2, and HGI-B1 vs. HGI-B2, are visualized by COVID19_GWAS_Analyzer. The top SNP displaying differential effect size between HGI-B1 and HGI-B2 is labeled as rs16831827.

The following codes further make local Manhattan plot (Figure 3) for target SNPs, including rs16831827 and other 7 SNPs that display association p value $< 1 \times 10^{-7}$ among any of these 3 GWASs. If user wants to add a gene track at the bottom of the local Manhattan plot, COVID19_GWAS_Analyzer provides another macro, called 'SNP_Local_Manhattan_With_GTF', to accomplish this task.

```
%Manhattan4DiffGWASs(
dsdin=D.GWAS1_vs_2,
pos_var=pos,
chr_var=chr,
P_var=GWAS1_P,
Other_P_vars=GWAS2_P Pval,
rm_signals_with_logP_lt=0,
uniq_colors=&_chr_colors_,
flip1stGWAS_signal=0,
sep_chr_grp=1,
fig_width=1200,
fig_height=600,
angle4xaxis_label=90,
xgrp_y_pos=-0.2,
yoffset_setting=%str(offset=(25,0.5)),
draw_local_Manhattan=1,
target_snps=rs16831827 rs13050728 rs13079478 rs14334143 rs2166172 rs2269899 rs622568
rs920065566
);
```

Figure 3 Local Manhattan plot for target SNPs among 3 GWASs

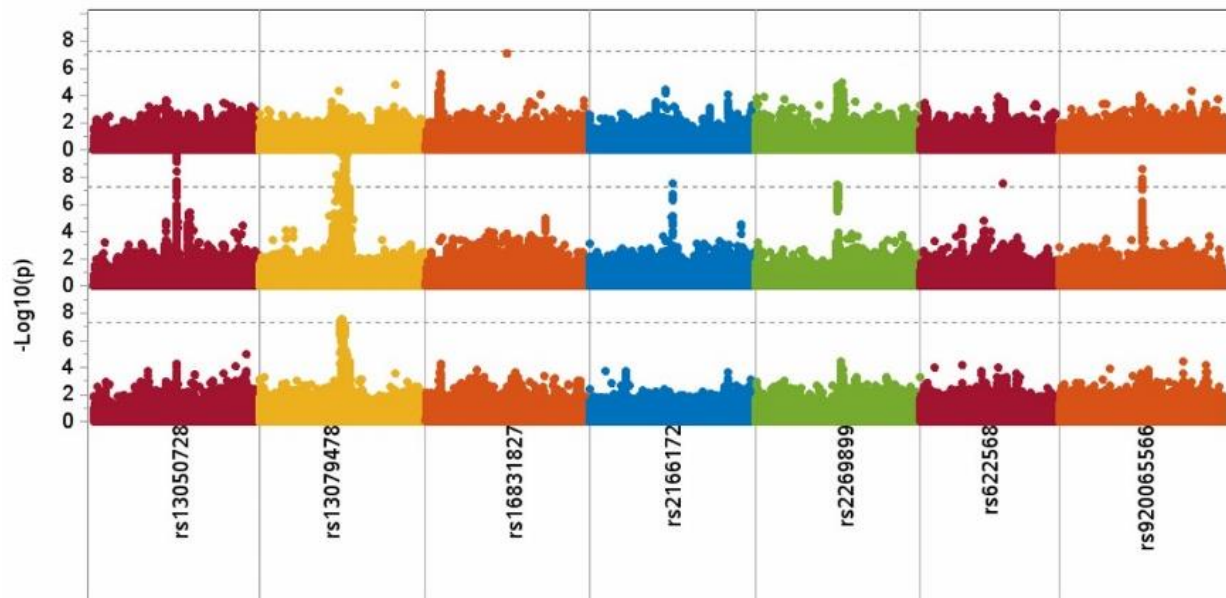


Figure 3. Local Manhattan plot for target SNPs among 3 GWASs. Target SNPs from three COVID-19 GWASs, including HGI-B1, HGI-B2, and HGI-B1 vs. HGI-B2 (from bottom to top), are visualized with local Manhattan plot by COVID19_GWAS_Analyzer.

EXPRESSION QUANTITATIVE TRAIT LOCI (EQTL) ANALYSIS

Many SNPs associated with disease risk are located in non-coding regions of the genome and may exert their effects by regulating gene expression. To assess whether the identified SNPs influence gene expression, we performed eQTL analysis using the following SAS macro:

```
%CaculateMulteQTLs_in_GTEx(
  query_snps=rs16831827,
  gene=MAP3K19,
  genoexp_outdsd=geno_exp,
  eQTLSumOutdsd=AssocSummary,
  create_eqtl_boxplots=1
);
```

This macro utilizes the GTEx dataset to evaluate the correlation between SNP genotype and gene expression levels across 49 tissues. For the SNP rs16831827, it is revealed a significant association with the expression of *MAP3K19* in lung tissue, suggesting a potential regulatory role for this variant in the context of COVID-19. The correlation is visualized through boxplots (Figure 4), which displayed the relationship between genotype and expression levels.

Figure 4 GTEx expression quantitative trait locus analysis for SNP rs16831827

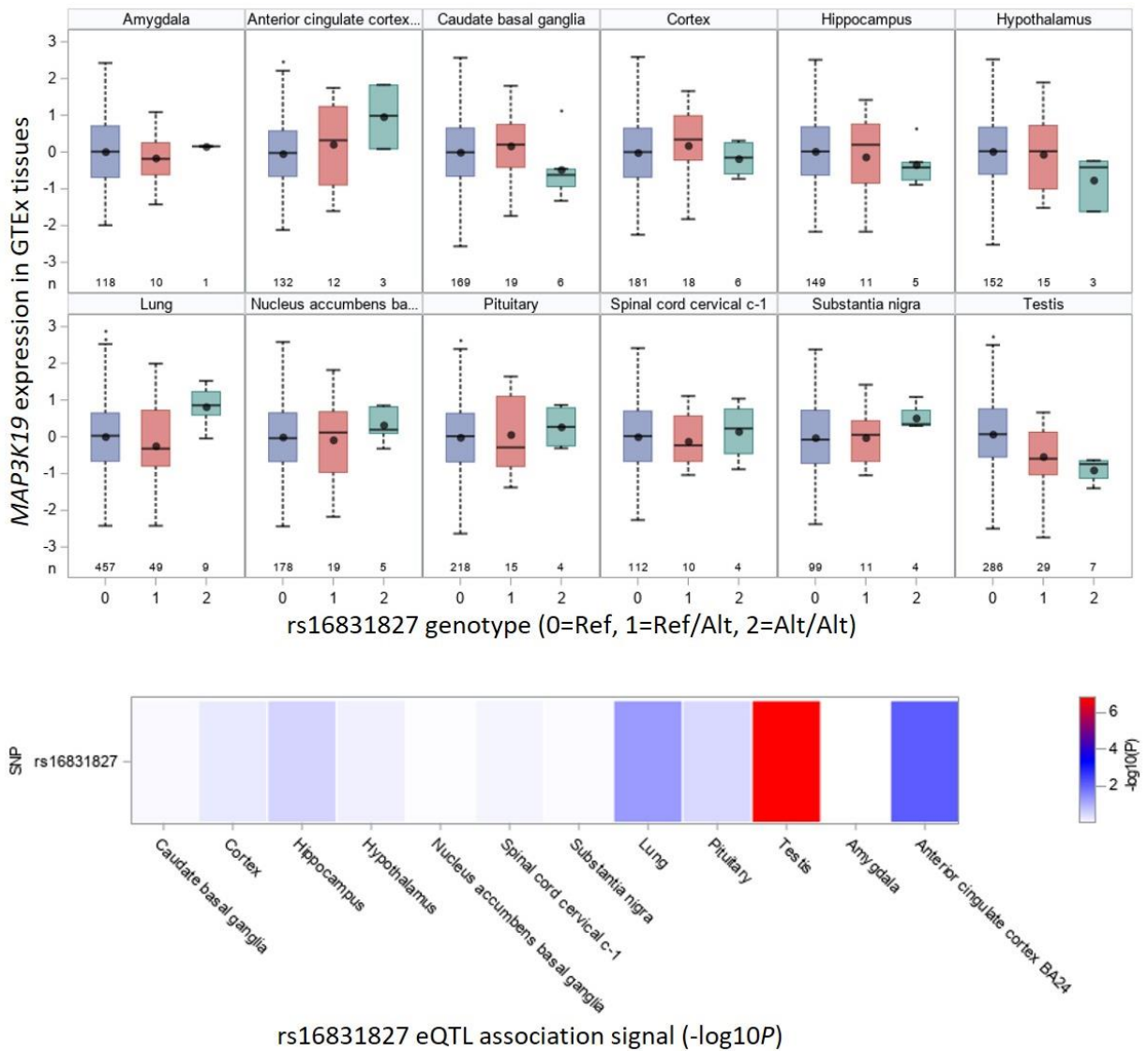


Figure 4. GTEx expression quantitative trait locus analysis for SNP rs16831827. *MAP3K19* is highly expressed in the genotype 'Alt/Alt' of rs16831827 in the lung tissue, which is the most relevant tissue to SARS-CoV2 infection.

SINGLE-CELL RNA EXPRESSION ANALYSIS

To further explore the biological relevance of key genetic variants, single-cell RNA-sequencing (scRNA-seq) data from nasopharyngeal samples of COVID-19 patients (Trump et al. 2021) is analyzed. The scRNA-seq analysis focuses on gene expression profiles in different cell types, resulting in the identification of the cell types most related to COVID-19 severity.

The following macro imports single-cell data and the visualization of *MAP3K19* gene expression among different COVID-19 patient groups by comparing with health controls:

```
%import_sc_mt看_meta_umap_data(
umap_file=https://cells.ucsc.edu/covid-hypertension/Seurat_umap.coords.tsv.gz,
```

```

meta_file=https://cells.ucsc.edu/covid-hypertension/meta.tsv,
cell_id_in_meta=cell,
exp_matrix_file=https://cells.ucsc.edu/covid-hypertension/exprMatrix.tsv.gz,
outdir=%sysfunc(pathname(HOME)),
/*Three sas data sets will be created and put into the specified HOME directory:
exp (read matrix with column headers, and the last column is for genesymbol),
umap (umap coordinates with sample meta data merged),
headers (cell barcodes corresponding to column headers for the exp matrix)*/
target_genes=MAP3K19,
/*Provide genesymbols separated by | for parsing lines match with these genes;
when the single cell data set is too large, this will save disk space in SAS OnDemand for Academics*/
mean_exp_cutoff=0.01, /*Only keep row records with mean read of gene expression >
mean_exp_cutoff*/
max_cells2import=1000000 /*Maximum number of cells to be imported;
if there are more than 1 million cells in the input cell matrix, only 1 million cells
will be randomly selected.*/
);

*Note: if the total number of cells in the matrix is more than 1 million, the macro will randomly select 1
million cells automatically, which will avoid using up the limited disk space (~5GB) in SAS OnDemand for
Academics.;

```

To draw single-cell UMAP (Uniform Manifold Approximation and Projection for Dimension Reduction) for the current single-cell RNA-seq dataset, the following command will generate the Figure 5 (shown at the end of this section).

```

libname sc "%sysfunc(pathname(HOME))";
%sc_umap(
umap_ds=sc.umap(where=(cluster not contains "Outlier")),
xvar=x,
yvar=y,
cluster_var=cluster,
fig_width=1000,
fig_height=1000
);

```

To visualize the gene expression of *MAP3K19* among different single-cell types, the following SAS macro “sc_scatter4gene” is used to generate single-cell UMAP and boxplots, illustrated in the Figure 6 at the end of this section.

```

libname sc "%sysfunc(pathname(HOME))";
*Modify phenotype categories;

```

```

data sc.umap;
set sc.umap;
if severity="control_healthy" then severity="Healthy";
if severity="severe" then severity="Severe";
if severity="critical" then severity="Critical";
run;
%sc_scatter4gene(
dsd=sc.exp,
dsd_headers=sc.headers,
dsd_umap=sc.umap,
gene=MAP3K19,
pheno_var=severity,
pheno_categories=Healthy Severe Critical,
boxplot_width=1000,
boxplot_height=600,
umap_width=1000,
umap_height=400,
umap_lattice_nrows=1,
boxplot_nrows=3,
where_cnd4sgplot=%quote(cluster not contains %'Outlier%')
);

```

Let's further zoom into specific UMAP area defined by the x- and y-axis for the single-cells mainly expressing *MAP3K19* by running the following codes (shown in Figure 7 at the end of this section):

```

*Note: the input SAS data set "new__tgt_dsd_" is internally generated by the above SAS macro;
%umap_with_axes_restriction(
dsdin=new__tgt_dsd_,
umap_width=800,
umap_height=300,
lattice_or_not=0,
raxis_max=20000,
raxis_min=0,
caxis_max=,
caxis_min=45000,
panel_row_num=1,
noheader=0
);

```

To better investigate these single-cell types expressing *MAP3K19*, the following codes are used to characterize the % of single-cell type with *MAP3K19* expression, # of each single-cell type with *MAP3K19* expression, sample-level *MAP3K19* expression, and single-cell level *MAP3K19* expression among health controls, severe COVID-19, and critical COVID-19 patients, as displayed in the Figure 8 at the end of this section.

```

*Use the dataset "new__tgt_dsd_" internally generated by running;
*the SAS macro sc_scatter4gene;
*The following steps combine non-ciliated cells as "Other" and
*treat samples with age>60 as "Yes" and others as "No".;
data tgt;
length cell_type $50.;
set new__tgt_dsd_;
cell_type="Other";
if prxmatch('/Ciliated/i',Cluster) then cell_type=Cluster;
if age>60 then old='Yes';
else Old='No';
run;
*Get percentage of different cells in the single cell data set;
*Note: we use the exp_cutoff=-1;
*This will calculate the percentages of different cells within each pheno group;
*percent=cells/total_cells_in_a_pheno_grp;
*The macro also perform differential gene expression analysis based on log10(reads+1)!;
%sc_freq_boxplot(
longformdsd=tgt,
cell_type_var=cell_type,
sample_grp_var=sample,
pheno_var=severity,
cust_pheno_order=Healthy Severe Critical,
exp_var=exp,
exp_cutoff=-1,
boxplot_height=250,
boxplot_width=800,
boxplot_nrows=1,
where_cnd_for_sgplot=%quote( cell_type contains 'Ciliated' or cell_type='Other'),
freqout=cellfreqout,
other_glm_classes=sex medication Old CAD CVD hypertension,/*covariant for the linear model*/
aggre_sc_glm_pdiff_dsd=all_sc);

```

Figure 5 Single-cell RNA-seq UMAP

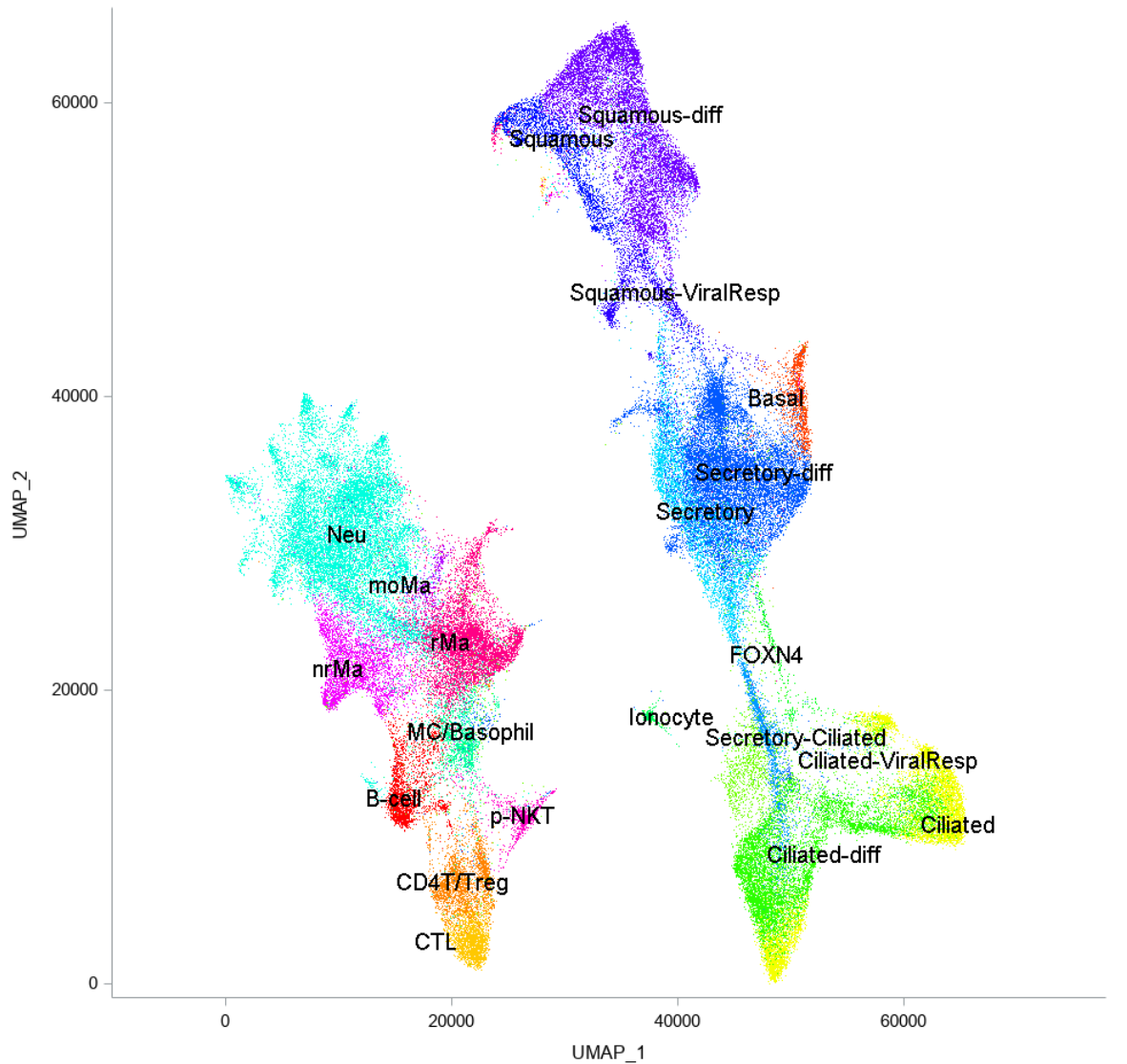


Figure 5. Single-cell RNA-seq UMAP. Different single-cell types are illustrated in the UMAP with a single SAS macro “sc_umap”. Note: names of single-cell types are labels in the UMAP.

Figure 6 UMAP and boxplots display *MAP3K19* mainly expressed in ciliated single-cells among health controls, severe, and critical COVID-19 patients.

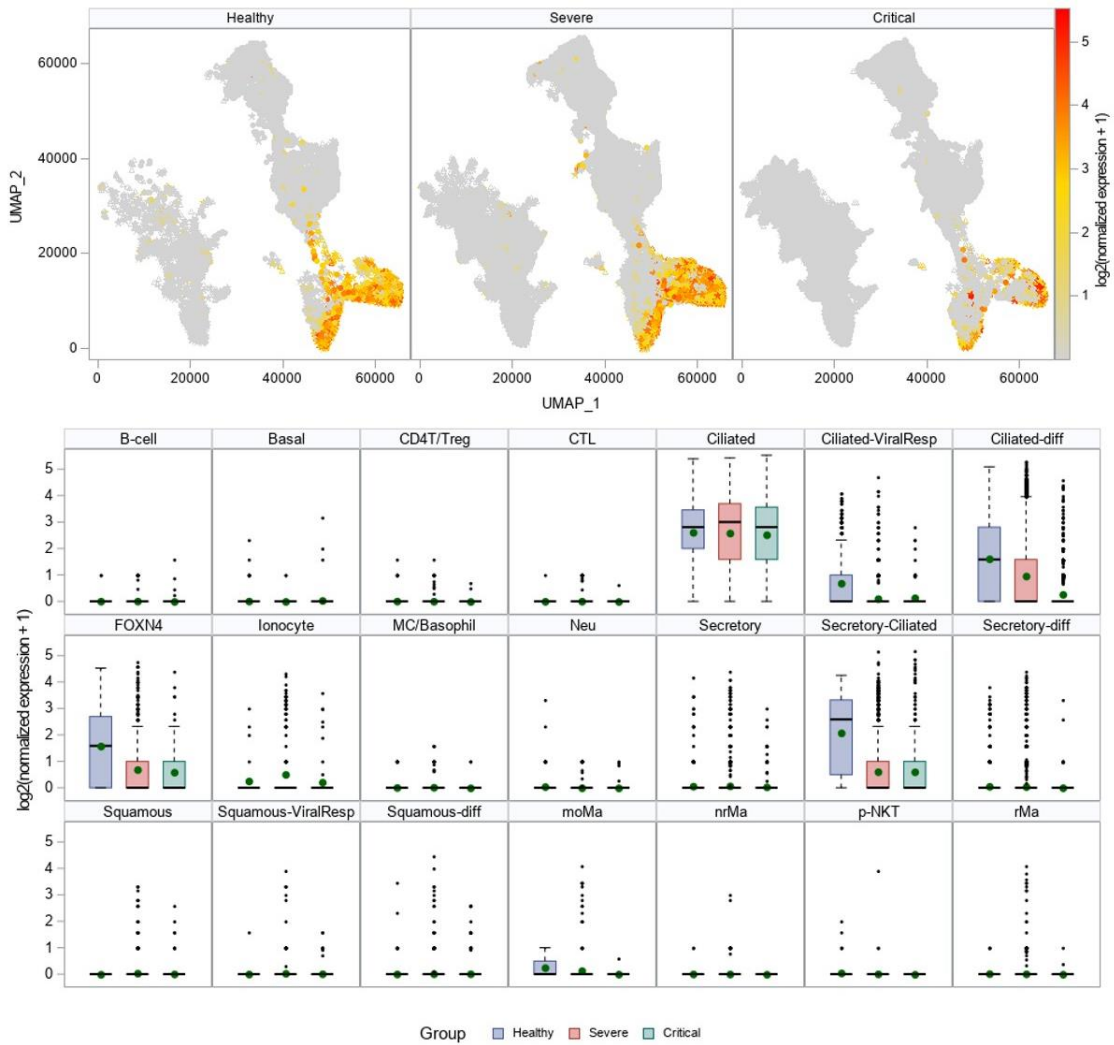


Figure 6. UMAP and boxplots display *MAP3K19* mainly expressed in ciliated single-cells among health controls, severe, and critical COVID-19 patients. The publicly available single-cell RNA-seq dataset published by Trump at al. (2021) is downloaded from UCSC Cell Browser (Speir at al. 2021) and then analyzed by COVID19_GWAS_Analyzer.

Figure 7 *MAP3K19* expression among ciliated single-cells

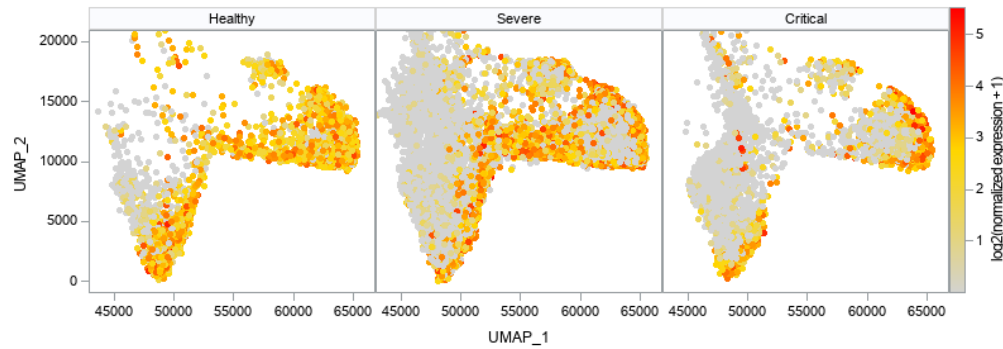


Figure 7. *MAP3K19* expression across ciliated single-cells. Single-cell RNA-seq UMAP illustrates the major expression of *MAP3K19* in the single-cell types of cilia and lower expression of *MAP3K19* in the critical COVID-19 patients compared to health controls and severe COVID-19 patients.

Figure 8 Descriptive statistics of single-cell types when evaluating gene expression of *MAP3K19*

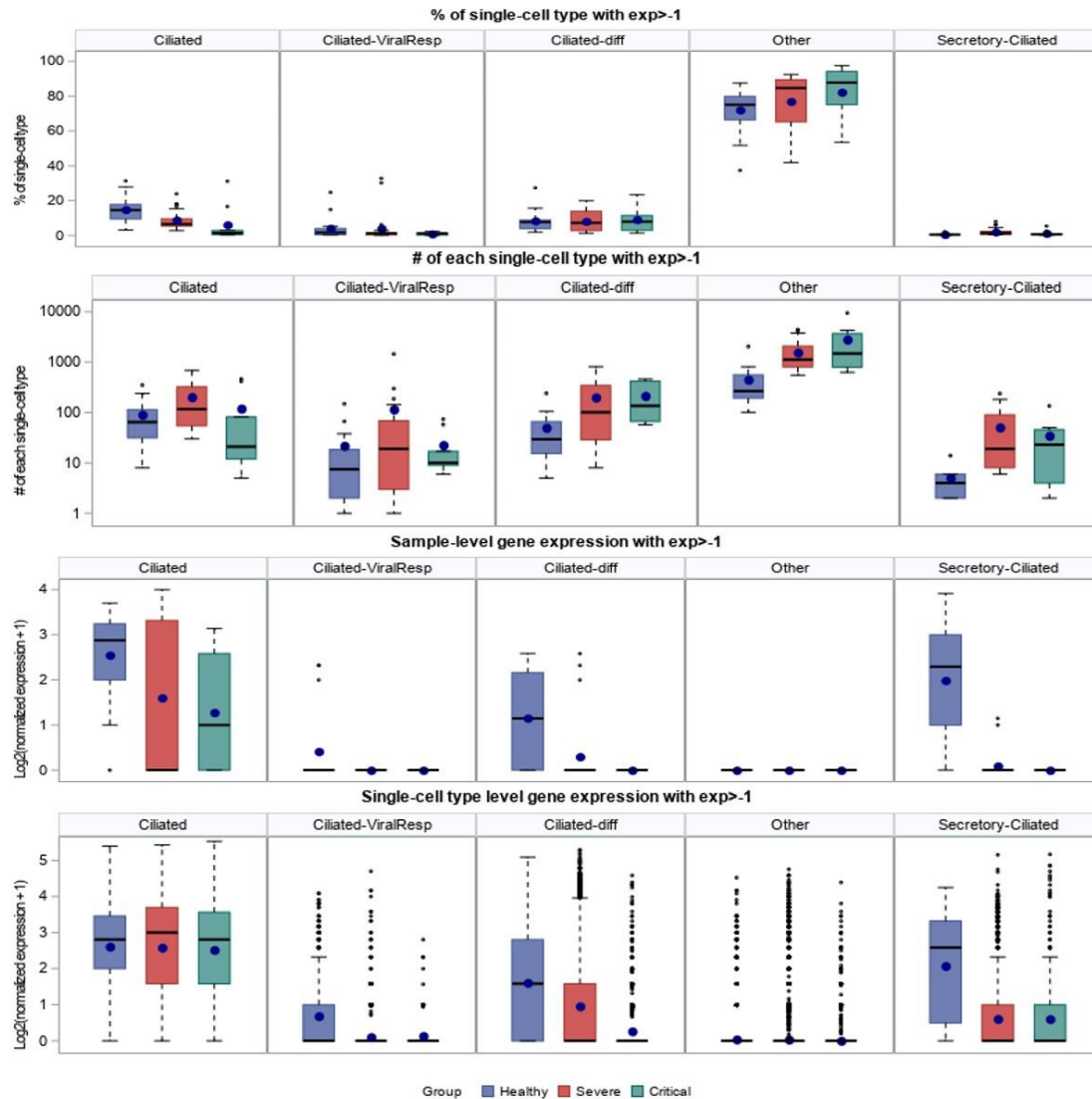


Figure 8. Descriptive statistics of single-cell types when evaluating gene expression of *MAP3K19*. *MAP3K19* expression across ciliated single-cells and other single-cell types are characterized for the % of single-cell type with *MAP3K19* expression, # of each single-cell type with *MAP3K19* expression, sample-level *MAP3K19* expression, and single-cell level *MAP3K19* expression. To include these single-cells that with expression value 0, the threshold of “exp>-1”, representing expression value of *MAP3K19* in a single-cell greater than -1, is employed to include them.

Taken together, the above single-cell RNA-seq analyses reveal that *MAP3K19* is primarily expressed in ciliated epithelial cells in the nasopharynx (Figure 5-8). These cells are related to the body's defense against respiratory infections, indicating that genetic variants influencing *MAP3K19* expression may alter the defense to SARS-CoV-2 infection.

CONCLUSION

The COVID19_GWAS_Analyzer provides a powerful and accessible platform for analyzing and visualizing GWAS data related to COVID-19 severity. By integrating differential GWAS comparisons, eQTL analysis, and single-cell RNA-seq data, this tool helps researchers uncover genetic variants that may explain the disparate impacts of COVID-19 across populations.

REFERENCES

- COVID-19 Host Genetics Initiative. 2021. “Mapping the human genetic architecture of COVID-19.” *Nature* 600, 472-477. 10.1038/s41586-021-03767-x.
- COVID-19 Host Genetics Initiative. 2022. “A first update on mapping the human genetic architecture of COVID-19.” *Nature* 608, E1-E10. 10.1038/s41586-022-04826-7.
- COVID-19 Host Genetics Initiative. 2023. A second update on mapping the human genetic architecture of COVID-19. *Nature* 621, E7-E26. 10.1038/s41586-023-06355-3.
- Speir, M.L., Bhaduri, A., Markov, N.S., Moreno, P., Nowakowski, T.J., Papatheodorou, I., Pollen, A.A., Raney, B.J., Seninge, L., Kent, W.J., and Haeussler, M. 2021. “UCSC Cell Browser: Visualize Your Single-Cell Data.” *Bioinformatics*. 10.1093/bioinformatics/btab503.
- Trump, S., Lukassen, S., Anker, M.S., Chua, R.L., Liebig, J., Thurmman, L., Corman, V.M., Binder, M., Loske, J., Klasa, C., et al. 2021. “Hypertension delays viral clearance and exacerbates airway hyperinflammation in patients with COVID-19.” *Nat Biotechnol* 39, 705-716. 10.1038/s41587-020-00796-1.
- Pairo-Castineira, E., Rawlik, K., Bretherick, A.D., Qi, T., Wu, Y., Nassiri, I., McConkey, G.A., Zechner, M., Klaric, L., Griffiths, F., et al. 2023. “GWAS and meta-analysis identifies 49 genetic variants underlying critical COVID-19.” *Nature* 617, 764-768. 10.1038/s41586-023-06034-3.
- Cheng, Z., Cai, Y., Zhang, K., Zhang, J., Gui, H., Luo, Y.S., Zhou, J., and DeVeale, B. 2023. “MAP3K19 regulatory variation in populations with African ancestry may increase COVID-19 severity.” *iScience* 26, 107555. 10.1016/j.isci.2023.107555.
- Zhang, K., Lin, S., Luo, Y.S., and Cheng, Z. 2024. “Protocol to search for genetic factors related to severe COVID-19 by analyzing publicly available genome-wide association studies.” *STAR protocols* 5, 103028. 10.1016/j.xpro.2024.103028.

ACKNOWLEDGMENTS

It is necessary to thank all researchers who generously share these COVID-19 related GWAS summary statistics, gene expression data from GTEx, and single-cell RNA-seq data.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Zhongshan Cheng
Center for Applied Bioinformatics, St. Jude Children's Research Hospital
Cheng.zhong.shan@gmail.com