Figure S1

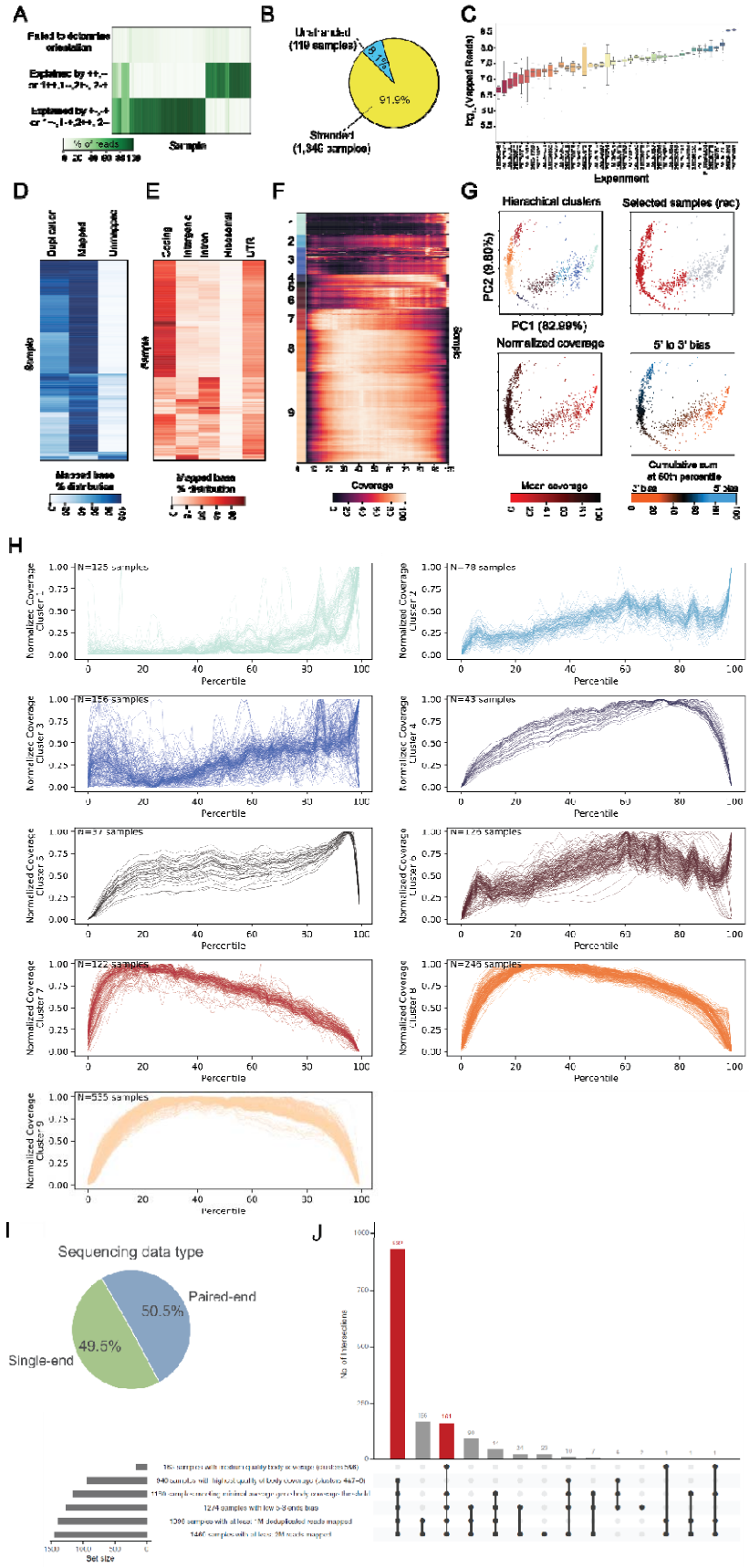

Figure S1.

A Heatmap of results from automatic detection of RNA-seq strandness.

B Pie chart shown percentage of stranded RNA-seq data versus unstranded.

C Boxplot of mapped reads number by studies.

D Heatmap of mapping rate, duplication rate.

E Heatmap of read genomic distribution.

F Hierarchical clustering of read aggregative gene-based coverage.

G PCA plot colored based on hierarchical clusters (in panel G), 5' to 3' bias score and Average normalized coverage. Number in brackets indicated variance explained.

H Line plot of aggregative gene-based coverage for individual cluster (defined in Supplementary Figure 1F).

I Pie chart shown the percentage of RNA-seq sample using single-end versus paired-end sequencing.

J Upset plot summaries the overlap between different clusters, groups and quality control criterions. The black dot in bottom right matrix indicated which group the sample for that bar belong to. For example, the 166 samples for second bar were overlapping only between group 1396_samples_with_at_least_1M_deduplicated_reads_mapped and 1460_samples_with_at_least_2M_reads_mapped.

Figure S2

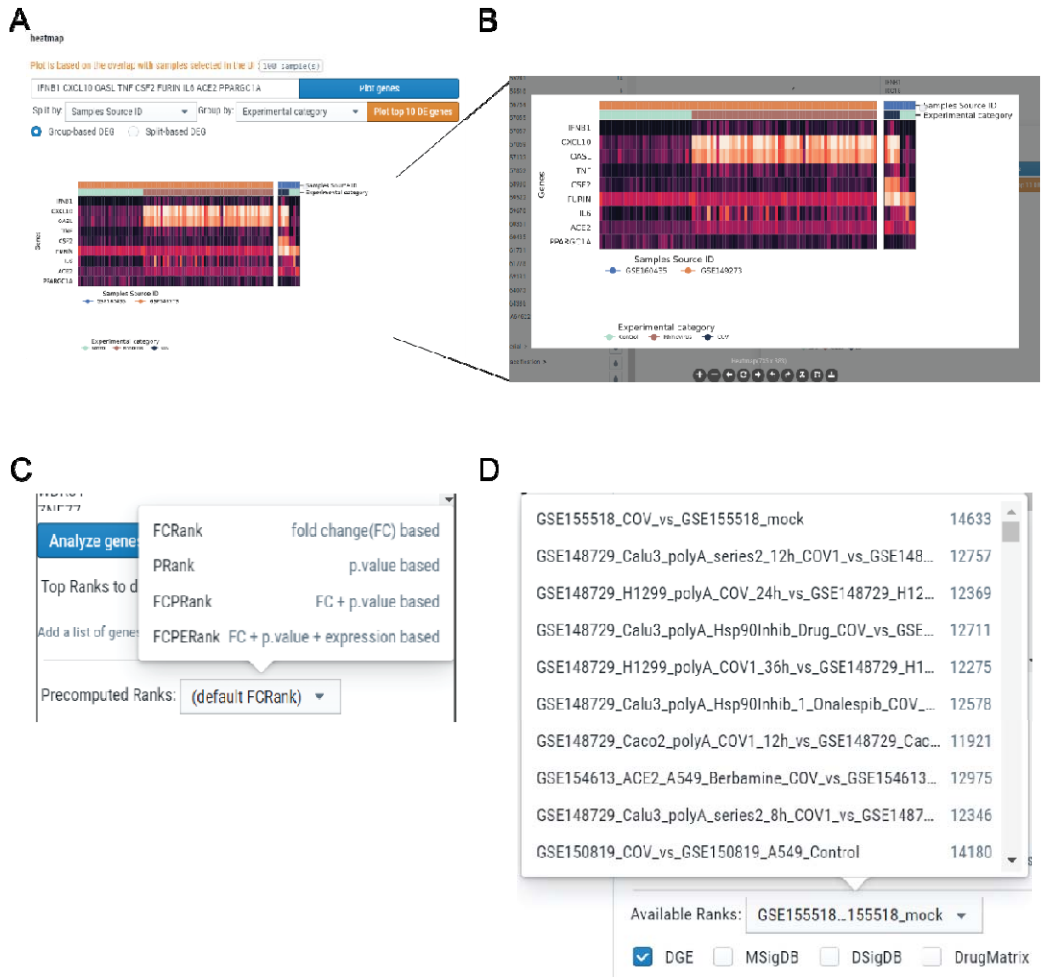

Figure S2.

A Detailed demonstration of plot function. These functions were made available for all types of plots as violin plot, dot plot, track plot and heatmap plot. User could input their gene list. User could also simple plot the top 10 differential expressed gene by selected meta data ("Source" as example) here. Results here indicated very few genes could be distinct one source from the others. For selected group 1 versus group2, differentially expressed genes between group 1 and group 2 could also been plot.

B Plot customization functions were provided so user could zoom, rotate, change text size, download the figure on-the-fly.

C We provide GSEA for different options of pre-computed ranks. User could choose based on different questions to ask. For example, using fold change (FC) based rank if they wondering what have been enriched by most changed genes. p.value based if they more interesting in what have been enriched by most consistently changed genes. See Material and Methods part for details.

D An example of how pre-computed results were organized and available for user's selection.

Figure S3

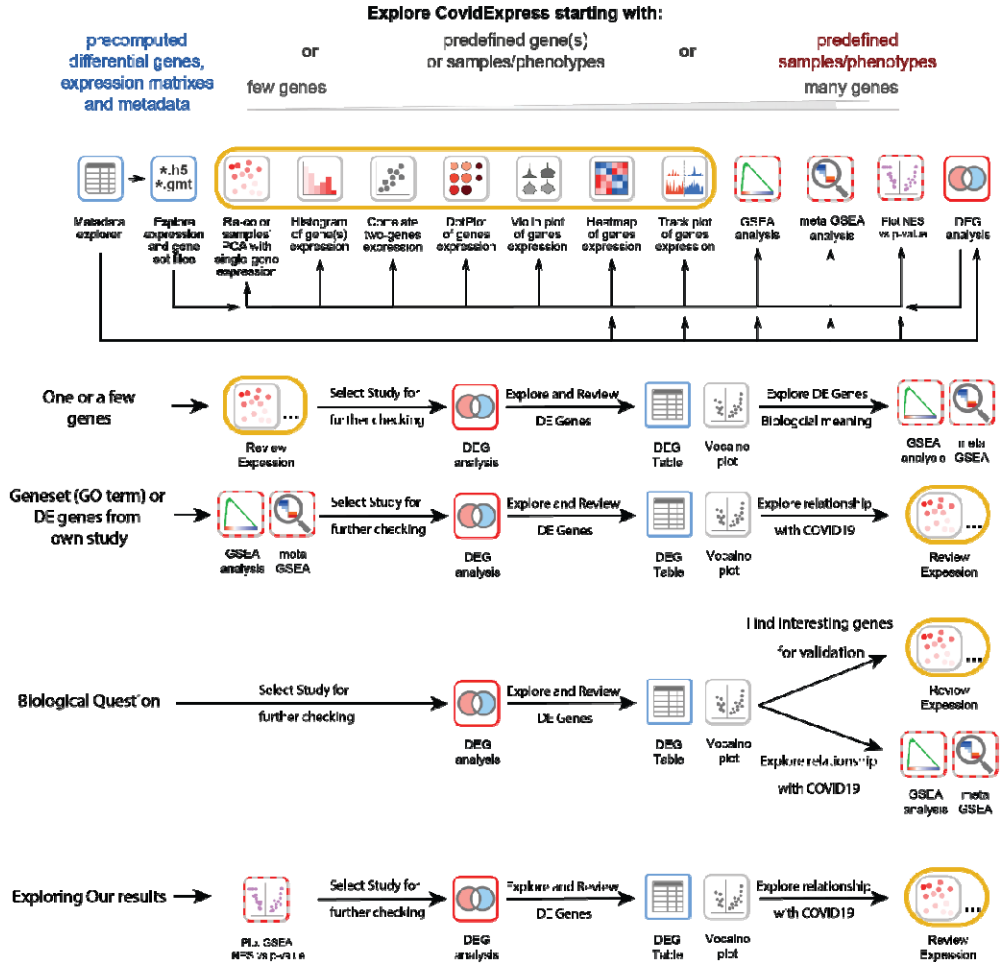

Figure S3.

Suggested analysis steps for variously investigation interests.

Figure S4

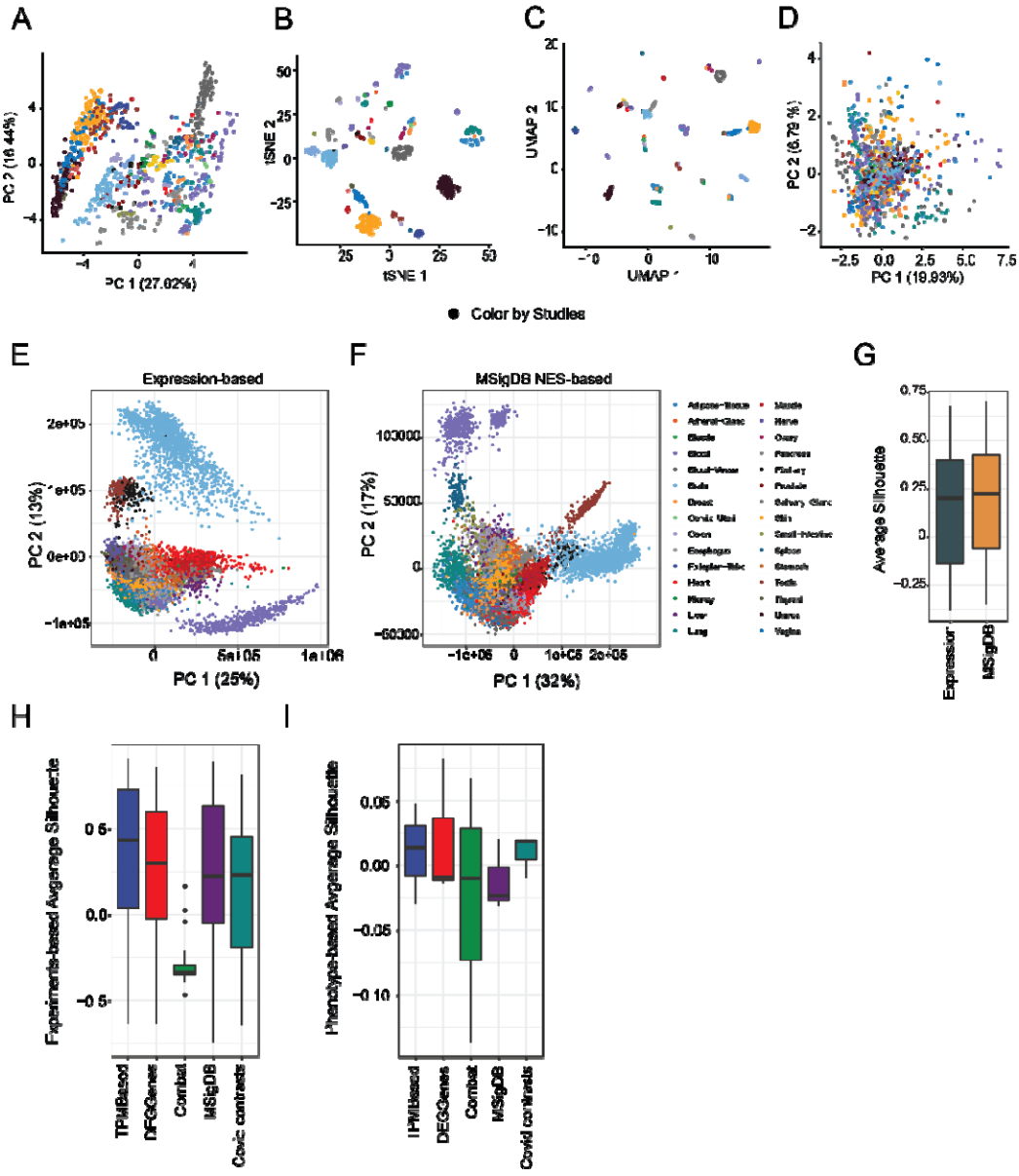

Figure S4.

A PCA analysis based on the top one thousand differential genes, colored by studies

B tSNE analysis based on gene expression level, colored by studies

C UMAP analysis based on gene expression level, colored by studies

D PCA analysis based on gene expression level after batch correction, colored by studies

E PCA analysis of GTEx data based on gene expression, colored by tissue.

F PCA analysis of GTEx data based on MsigDB signature, colored by tissue.

G Comparison of tissue-level Silhouette score distribution using expression and MSigDB-based PCA projections.

H Experiment-level Silhouette scores distribution between our compiled samples using different scoring methods to measure batch-effect. Lower score indicates less batch effect.

I Phenotype-level Silhouette scores distribution between our compiled samples using different scoring methods to measure the degree of phenotype separability. Higher scores indicate better separability between SARS-CoV-2 infection and control samples.

Figure S5

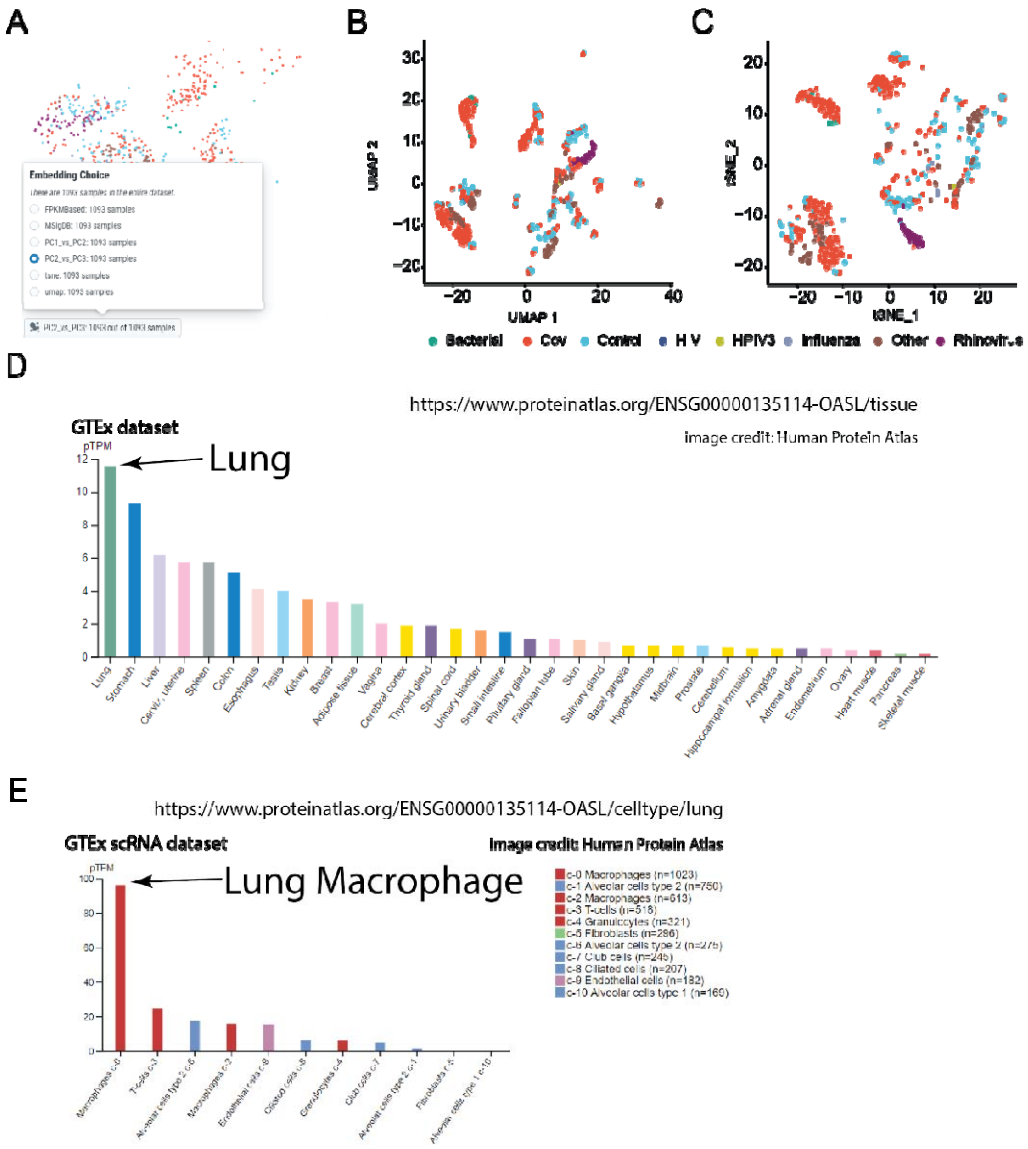

Figure S5.

A An example shown various choices of samples embedding layouts user could use.

B UMAP analysis based on single-sample Gene Set Enrichment Analysis(ssGSEA), using COVID signature gene sets from this study, colored by SARS-CoV-2 infection status.

C tSNE analysis based on single-sample Gene Set Enrichment Analysis(ssGSEA), using COVID signature gene sets from this study, colored by SARS-CoV-2 infection status.

D *OASL* expression in different tissues from GTEx datasets.

E *OASL* expression in different cell types from Lung GTEx single-cell RNA-seq datasets.