Download of the datasets We downloaded the data sets GSE156063, GSE163151 related to COVID-19 and GSE2411, GSE17355 related to ALI from the Gene Expression Omnibus (GEO) database through the R package "GEOquery". GSE156063 is from *Homo Sapiens* with GPL24676 data platform, containing 193 upper respiratory tract cell samples from 93 COVID-19 patients and 100 healthy people, all included in this study. GSE163151 is also from *Homo Sapiens*, with GPL24034 data platform, containing 176 nasal swabs (upper respiratory tract cells) from 145 COVID-19 patients and 31 healthy people, all included in this study. GSE2411 is from *Mus Musculus* with GPL339 data platform. We selected 6 ALI mice and 6 control mice in this study. GSE17355 is from *Mus Musculus*, and the data platform is GPL4865. We selected the whole lung samples from 9 ALI mice and 3 control mice into this study. The GPL platform file corresponding to the chip was used in the probe name annotation. Check Supplementary table 1 for more information.

Differential expression analysis We first used R packet "sva" to remove batch effects on the COVID-19 datasets and ALI datasets, respectively. We merged the two datasets into one COVID-19 dataset and compared the datasets before and after removing batch effects by using distribution box plots. The same operation was done for the ALI dataset. Our approach utilized ComBat_seq (from the R sva package v3.46.0), a negative binomial model specifically designed for RNA-seq count data, to adjust for technical variations across sequencing batches while preserving biological group effects. We conducted differential analysis between the disease group and the control group on both the datasets and selected differentially expressed genes (DEGs) by using $|logFC|>0$ and $P.adj<0.01$ as the threshold.

Gene set variation analysis (GSVA) GSVA aggregates the expression levels of genes

into a comprehensive score for a specific pathway, and then evaluates the changes in this score under different conditions. We obtained the gene set from MSigDB and performed GSVA analysis on the datasets to obtain the enriched pathway with the screening criterion as $P<0.05$.

Single sample GSEA (ssGSEA) algorithm The enrichment score calculated through ssGSEA in "GSVA" package is performed to represent the ssGSEA enrichment score for 28 cells in each sample. The differences in ssGSEA enrichment score between the disease and the control group (or high-score and low-score group) in the datasets of 28 types of immune cells were displayed with boxplots, and the correlations between immune cells and hub genes were displayed through correlation heatmaps.

Feature gene selection by LASSO The least absolute shrinkage and selection operator (LASSO) regression based on linear regression by adding penalty terms for better fitting and maximizing the generalization ability of the algorithm. We used 10x cross validation on the datasets with "seed number" as "123", and "cycles" as "1000" to perform the analysis. The results (feature genes) of the two datasets were intersected as key genes for subsequent analysis. In addition, we included only the disease samples from the datasets (COVID-19 dataset, ALI dataset) and used the median risk score to divide them into high-score and low-score groups.