Alterations in the host transcriptome in vitro and in vivo following severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection

Xiaomei Lei [1 #], Zhijun Feng [2 #], Xiaojun Wang [1*], Xiaodong He [2 *]

[1] Gansu Provincial Hospital. No. 204, Donggang West Road, Chengguan District, Lanzhou, Gansu, China.

[2] Lanzhou University Second Hospital. No. 82, Cuiyingmen, Chengguan District, Lanzhou, Gansu, China.

[#] These authors are contributed equally to this work.

[*] Corresponding author:

Xiaojun Wang

No. 204, Donggang West Road, Chengguan District, Lanzhou, Gansu, China.

E-mail address: wangxj19@lzu.edu.cn

Xiaodong He

No. 82, Cuiyingmen, Chengguan District, Lanzhou, Gansu, 730030, China.

E-mail address: hxd@lzu.edu.cn

1. Data processing information

GSE148815 is a dataset to analyze the cell-intrinsic differences associated with SARS-CoV-2 infection between between laser-captured fresh epithelial cells from children (6 samples) and adults (6 samples). GSE150316 is a total RNA-seq analysis of lung tissues devoid of acute inflammation of 5 patients deceased due to SARS-CoV-2 infection. GSE147507 is a total RNA-seq for evaluating the transcriptional response to SARS-CoV-2 infection. Three microarray data (patient-data in vivo level, and two cell lines data in vitro level) were conducted and performed data processing as described in the main text. The detailed steps processed in R software were listed as follows (Italics represent R code in data processing procedure):

1.1. Set up comparison subjects

sample = c (rep ('Healthy', sample number), rep ('Infection', sample number))
metadata$sample = relevel (factor (sample), " Healthy ")

1.2. Standardized microarray data

```
library(DESeq2)
dds = DESeqDataSetFromMatrix(countData=expr_df,
                             colData=metadata,
                             design=~sample,
                             tidy=TRUE)
```

1.3. Excluded genes with overall expression less than 1.

dds = dds[rowSums(counts(dds))>1,]

1.4. Continue to complete standardized process

vsd = vst(dds, blind = FALSE)
dds = DESeq(dds)

1.5. Extracted the data of gene expression after standardizing

normalized_counts = as.data.frame (counts (dds, normalized=TRUE))

1.6. log2 transformed

ex = normalized_counts

```
qx = as.numeric(quantile(ex, c(0., 0.25, 0.5, 0.75, 0.99, 1.0), na.rm=T))
LogC = (qx[5] > 100) ||
   (qx[6]-qx[1] > 50 && qx[2] > 0) ||
   (qx[2] > 0 && qx[2] < 1 && qx[4] > 1 && qx[4] < 2)

if (LogC) {
   ex[which(ex <= 0,arr.ind = T)] = NaN
   exprSet = log2(ex)
   print ("log2 transform finished")
} else {
   Print ("log2 transform not needed")
}
```

1.7. Handled missing or unknown values.

```
exprSet[is.na(exprSet)] = 0
```

1.8. Checked and adjusted batch effects.

Checking:

```
dist_mat = dist (t(edata))
clustering = hclust(dist_mat)
plot (clustering, labels = rownames(pheno), hang = -1)
```

Adjusting:

```
pheno$hasInfection = pheno$group == "Infection"
model = model.matrix(~hasInfection, data=pheno)
library(sva)
combat_edata = ComBat(dat = edata, batch = pheno$batch, mod = model)
dist_mat_combat = dist(t(combat_edata))
clustering_combat = hclust(dist_mat_combat, method = "complete")
plot (clustering_combat, labels = rownames(pheno),hang=-1)
```

1.9. Corrected logFC

```
contrast = c ("sample", "InfecCase", "NegControl")
dd1 = results (dds, contrast=contrast, alpha = 0.05)
plotMA (dd1, ylim=c(-2,2))
dd2 = lfcShrink (dds, contrast=contrast,type = "ashr", res=dd1)
plotMA (dd2, ylim=c(-2,2))
summary (dd2, alpha = 0.05)
```

1.10. Performed differential analysis

```
library(dplyr)
library(tibble)
res = dd2 %>%
```

data.frame() %>%
rownames_to_column("gene_id")

2. Results

2.1 Details of the microarray data

Data type	Samples		Platform
	SARS-CoV-2 infected	**SARS-CoV-2 un-infected**	
Patient-data	GSM4483226/GSM4483227/ GSM4483228/GSM4483229/ GSM4483230/GSM4483231	GSM4546608/GSM4546609/ GSM4546610/GSM4546611/ GSM4546612	GPL18573
NHBE-data	GSM4432378/GSM4432379/ GSM4432380	GSM4432381/GSM4432382/ GSM4432383	GPL18573
A549-data	GSM4432384/GSM4432385/ GSM4432386/GSM4462336/ GSM4462337/GSM4462338	GSM4432387/GSM4432388/ GSM4432389/GSM4462339/ GSM4462340/GSM4462341	GPL18573

Table s1 The details of the microarray data used in the study. No., number. NHBE, normal human bronchial epithelial. A549, lung adenocarcinoma cell line A549.

2.2. Batch effects

No batch effects were found in the patient-data and NHBE-data, as shown in Figure s1a and Figure s1b. A significant batch effect was found in A549-data (Figure s1c), and the adjusted result was shown in Figure s1d.

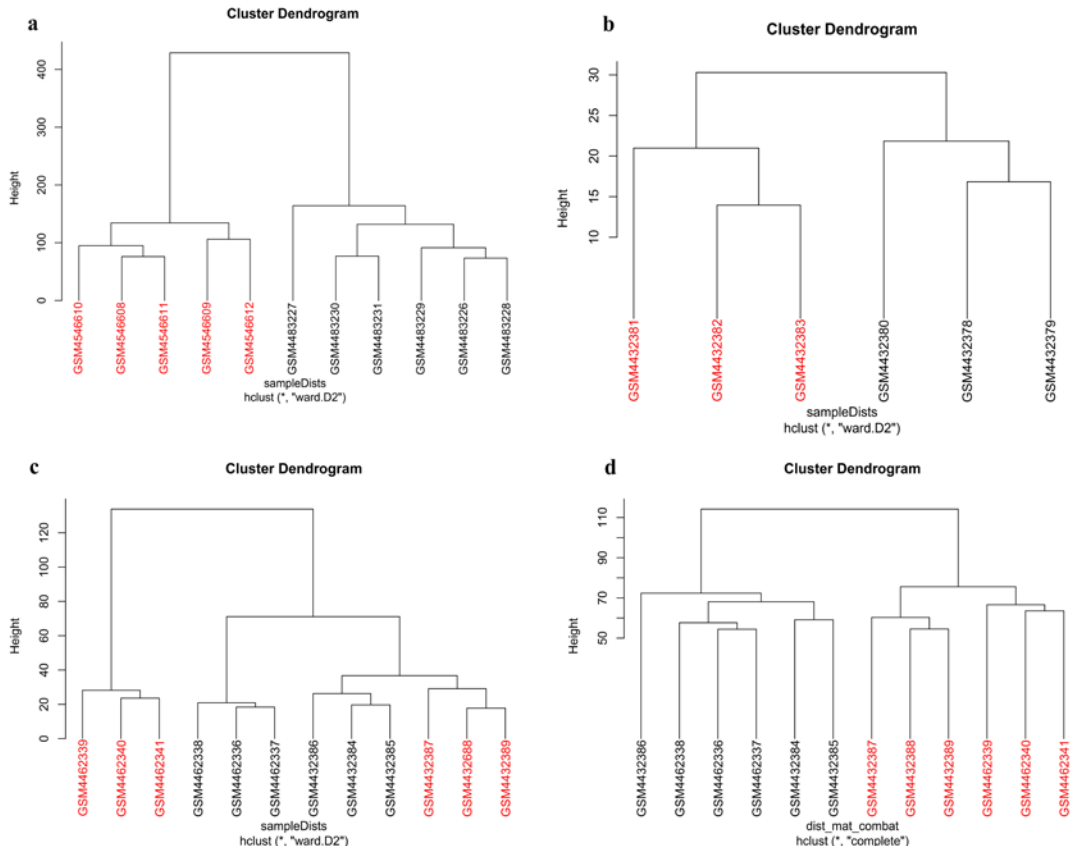

Figure s1 Batch effects in the three microarray data. a is for patient-data, b is for NHBE-data, c is for A549-data, and d is adjusted result for A549-data. The GSM number in red front represents SARS-CoV-2 infected samples while black represents normal healthy samples.

2.3. Corrected logFC

The original and the corrected results for the three data were shown in Figure s2a to Figure s2c.

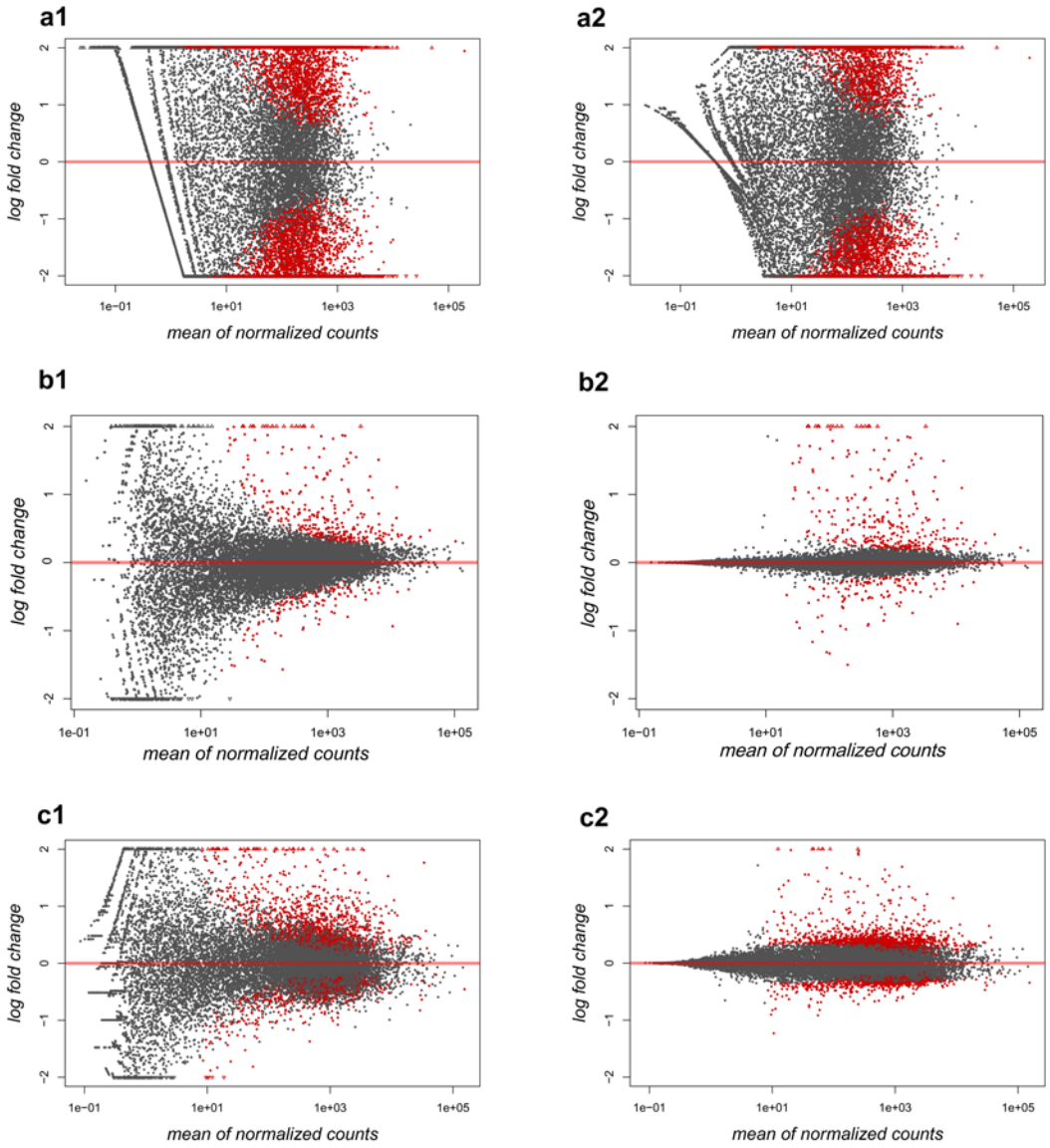

Figure s2 The original and the corrected logFC results for the three data. a1 is the original result of logFC in patient-data, a2 is corrected result for patient-data; b1 is original for NHBE-data, and b2 is corrected for NHBE-data; c1 is original for A549-data, and c2 is corrected for A549-data. Red points represent genes with |logFC|>0.5 and expression level more than 10.

2.4. Overall gene expression of each sample

After completing data processing, we visualized the overall gene expression of each sample in the three data, as shown in Figure s3a-c, which reflected that the gene expression was uniformly distributed in our study.

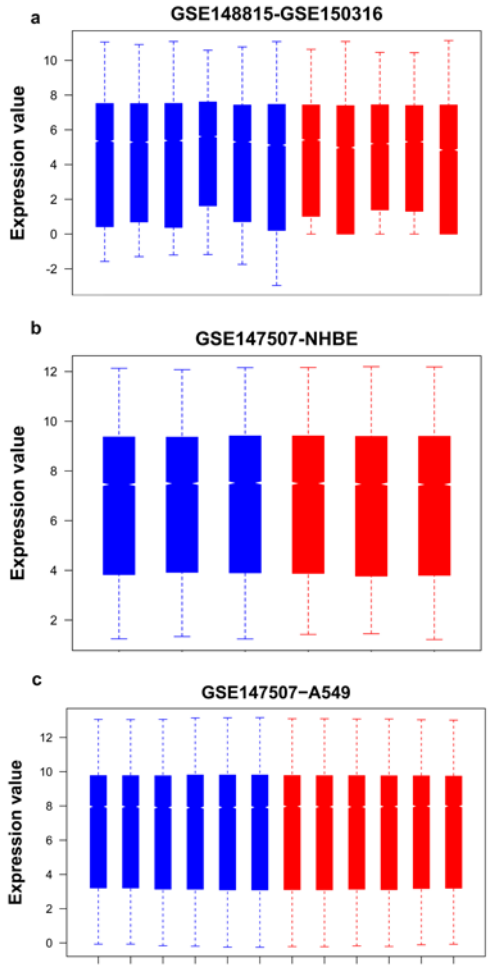

Figure s3 Overall gene expression of each sample. a is for patient-data; b is for NHBE-data, and c is for NHBE-data. Red represent SARS-CoV-2 infected samples, blue represent normal healthy samples.