Article

SARS-CoV-2 Infections - Gene Expression Omnibus (GEO) Data Mining, Pathway Enrichment Analysis, and Prediction of Repurposable Drugs/Compounds

Srilakshmi Chaparala, Carrie L Iwema and Ansuman Chattopadhyay *

Health Sciences Library System, University of Pittsburgh, Pittsburgh, PA, USA
* Correspondence: ansuman@pitt.edu

Supplementary Information:

Methods

SM1—GEO datasets

GSE147507 contains bulk RNA-Seq data from both virus-infected and mock-infected experiments performed on a variety of cell lines and animal models. Out of 110 samples, we selected data from three cell lines for our analysis: Normal Human Bronchial Epithelial (NHBE), A549—derived from explanted cultures of human lung cancer tissue, and Calu3—epithelial cells derived from lung adenocarcinoma. The three virus strains used for infection were: SARS-CoV-2 (USA-WA1/2020) with Multiplicity of Infection (MOI) 0.2 or 2.0, RSV (A2 strain; MOI 2.0), and human parainfluenza virus-3 (HPIV-3; MOI 3. 0). PolyA RNA was harvested 24 hours post-infection for SARS-CoV-2, RSV, and HPIV-3. GSE152075 contains RNA-Seq profiles of nasopharyngeal swabs from 430 individuals with SARS-CoV-2 and 54 negative controls.

SM2—CLC Genomics data

Raw reads for all human samples (GSE147507; series 1-9, 15, 16) were downloaded from the NCBI Sequence Read Archive [59] and data analysis was performed using CLC Genomics Workbench v20 (QIAGEN Digital Insights). Raw reads were aligned to the human reference genome (hg38) with annotations (Ensemble v91) using default settings, and unmapped reads were collected and aligned against corresponding virus reference genome sequence. The differential expression of genes between virus-infected vs mock-treated samples were computed for each series. Genes with FDR-adjusted p-value <0.05 and log fold change +/- 1.0 were considered DE genes.

SM3—BaseSpace Correlation Engine drug-related datasets

Dactinomycin: results from a microarray experiment in which A549 human lung cancer cells were seeded eight days before drug treatment (GSE6400) [60]. At four hours before RNA isolation, Dactinomycin (5 μg/mL final concentration) or control (5% mannitol) solution was added to the cell cultures.

Triptolide: results from a microarray experiment describing identification of genes affected by a short treatment of A549 cells with increased concentrations of Triptolide (GSE16760) [61]. The dataset includes DE genes from Triptolide-treated vs. untreated samples in which three concentrations (0.05, 0.15, 0.45 μM) of the drug were introduced at three treatment time points (1, 2, 4 hours).

Dexamethasone: results from four GEO datasets measuring gene expression in human lung tissue-derived cells treated with Dexamethasone. GSE34313 is a microarray dataset of airway smooth muscle cell responses to Dexamethasone at 4 and 24 hours [62]. GSE52778 is an RNA-Seq dataset after 18 hours of

Dexamethasone treatment [63]. GSE96649 is an RNA-seq dataset from A549 cells with Dexamethasone (1uM) for treatment at 5 hours vs. untreated controls [64] . GSE17307 is a microarray dataset of Dexamethasone (1uM) treatment at 6 hours vs. untreated controls [65].

Hydroxychloroquine: GSE74235 is an RNA-Seq dataset of human Peripheral Blood Mononuclear Cells from three healthy participants, stimulated with rheumatogenic, heat-killed Group A Streptococcus (GAS) for 24 hours, MOI 10 [66] . The effect of Hydroxychloroquine (20 μM) was measured in combination with GAS.

SM4—BaseSpace Correlation Engine pathway comparison

BSCE was used to mine the GEO datasets and compute DE genes between two conditions. For microarray experiments, statistical significance was determined using the Welch t-test. Genes with p-value cutoff 0.05 and fold change cutoff +/- 1.20 were considered DE genes. For RNA-Seq experiments, raw sequencing reads were aligned to the human genome sequence using STAR 2.3 and RefSeq annotations. Differential gene expression results between the control and test sample groups were calculated using DESeq2. Genes with Benjamini-Hochberg adjusted q-value cutoff 0.05 and a fold change cutoff of +/-1.2 were considered DE. BSCE software uses a ranked-based non-parametric Running Fisher test to compare gene expression signatures between two DE gene lists and calculate a correlation score [27].

Documents

SD1. IPA canonical pathway enrichment analysis results for DE genes from SARS-CoV-2-infected vs. mock control from series 16 (GSE147507). A spreadsheet containing pathway names, p-values, activity z-scores, and overlapping genes. DOI: 10.6084/m9.figshare.12899387

SD2. BSCE pathway enrichment analysis results for DE genes from SARS-CoV-2-infected vs. mock control from series 16 (GSE147507). A spreadsheet containing pathway names, pathway direction, and p-values. DOI: 10.6084/m9.figshare.12899399

SD3. IPA canonical pathway comparison across multiple samples. A spreadsheet containing pathway names and activity z-scores. DOI: 10.6084/m9.figshare.12925724

SD4. Predicted COVID-19 repurposable drugs/compound. The full list of BaseSpace Correlation Engine computed drugs or compound revealing a negative correlation with the gene expression profile of series16_GSE147507 (differentially expressed genes from A549 cells transfected with exogenous ACE2, infected with SARS-CoV-2 (USA-WA1/2020) vs. mock control). DOI: 10.6084/m9.figshare.12899408

Table

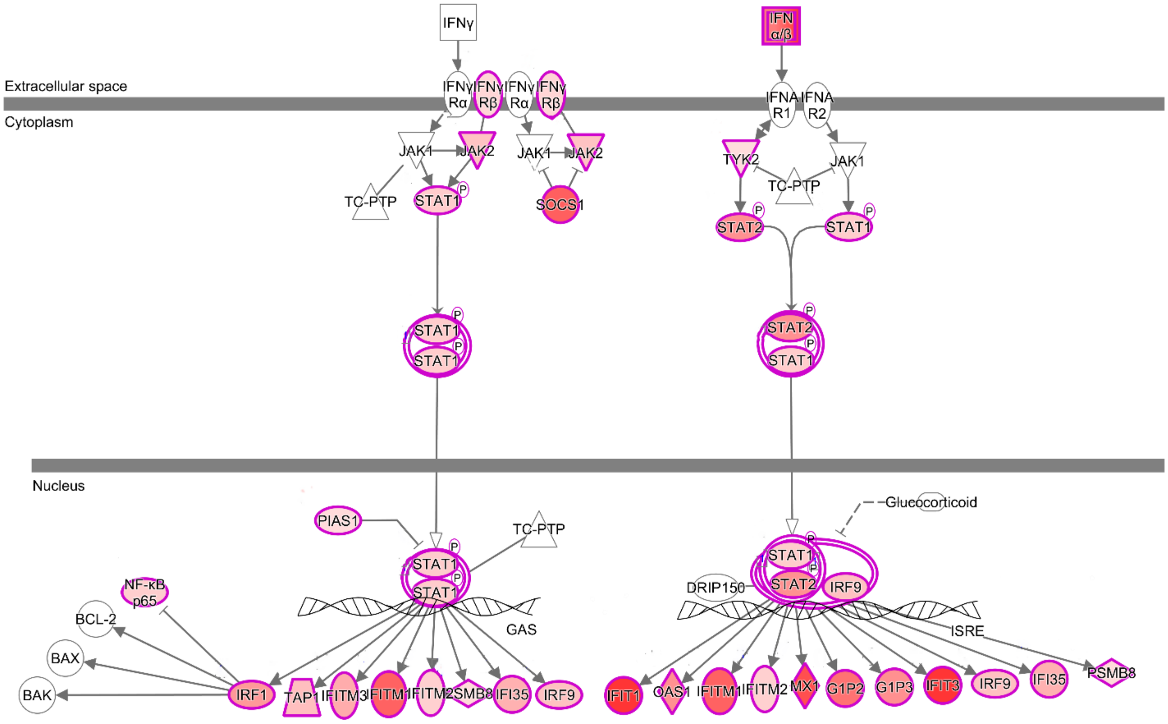
ST1. RNA-Seq analysis results (GSE147507) showing SARS-CoV-2 viral gene expression.

Gene	S1_SARSCOV2_TPM Mean	S2_SARSCOV2_TPM Mean	S5_SARSCOV2_TPM Mean	S6_SARSCOV2_TPM Mean	S7_SARSCOV2_TPM Mean	S16_SARSCOV2_TPM Mean
ORF1ab_1	0	0	0	5847.45	1413.60	5146.64
ORF1ab_2	1734.47	1820.60	4256.43	5782.00	832.73	1981.77
S	14069.43	29658.22	34891.84	19707.84	7974.92	12356.82
ORF3a	13245.19	53848.57	58476.78	35125.22	19570.48	28092.09
E	1086.75	19383.58	36610.71	15730.09	9225.75	14808.52
M	201586.32	119051.8	112120.11	99627.36	86629.40	73195.24
ORF6	3407.41	10348.67	12259.08	32704.57	10697.02	4592.52
ORF7a	24082.43	113181.1	118683.19	103757.22	95575.52	122842.32
ORF7b	0	315.0067	3462.13	5360.12	1578.68	580.01
ORF8	23198.55	105271.3	146962.49	117131.70	131925.11	155547.37

N	714286.99	545871.2	471327.45	519120.58	632843.03	539498.95
ORF10	3302.41	1250.03	949.80	40105.84	1733.76	41357.76

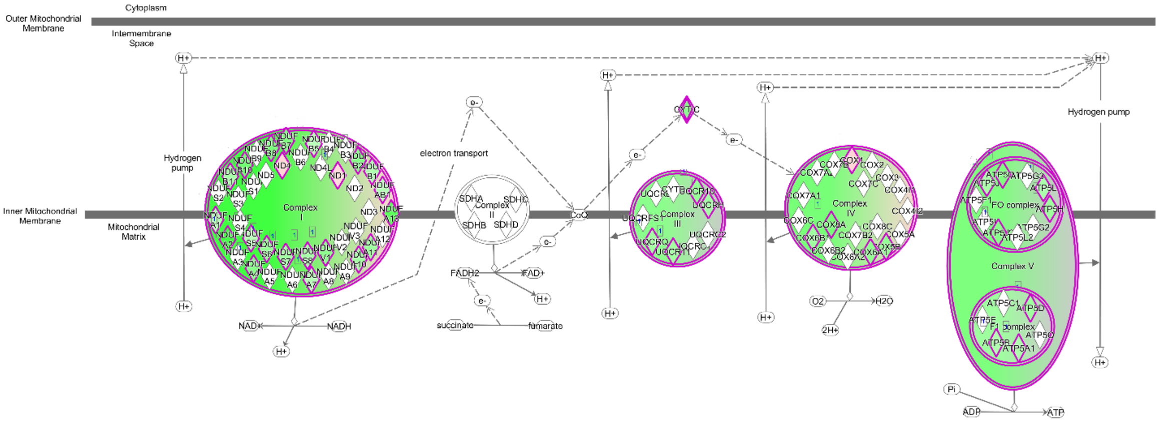
Figures

SF1. Interferon signaling pathway overlaid with differentially expressed genes observed between SARS-CoV-2-infected vs. mock control.



The fuchsia color indicated up-regulated genes with $\log_2$Fold change > 1. The intensity of the color is proportional to the magnitude of fold change.

SF2. Mitochondrial oxidative phosphorylation pathway overlaid with differentially expressed genes observed between SARS-CoV-2-infected vs. mock control.



The green color indicates down-regulated genes with $\log_2$Fold change < -1. The fuchsia color indicates up-regulated genes with $\log_2$Fold change > +1. The intensity of the color is proportional to the magnitude of fold change.