Re-analysis of Covid-19 related RNASeq count data reveals a robust list of genes that exhibit significant Covid-19 dependent differential expression

Kevin P. Keegan, Ph.D.[1*]

[1]Independent researcher, Glen Ellyn, IL, USA

* Corresponding Author

E-mail: kevin.p.keegan@gmail.com

Abstract

In recent years, for obvious reasons, a keen interest in discovering genes and pathways affected by the affliction of Covid-19 have led to numerous large scale studies utilizing RNASeq technology. These studies have led to a wide breadth of discovery related to understanding, and ultimately treatment of this dreaded disease. Notably, the results of many such studies have been deposited with the Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) and are freely available to the public. Here a novel analysis of the data produced from three such studies is presented. After a total of 4031 samples were screened for presence of unambiguous Covid-19 state related metadata, and some light filtering for quality, 2678 samples were used. The aim of this analysis was two-fold. First, to produce a robust list of genes that exhibit highly significant Covid-19 dependent differential expression across all three studies, and are likely to be of interest to the scientific and medical communities. Second, to provide the reproducible results of this study as an example for further investigation. The analysis includes solutions for procedural issues encountered - from download of data and metadata, to quality screening of samples, dimension reduction and visualization of data from all samples in each study, to a pathway analysis of the final results, a list of some 61 genes (60 annotated) found in common among all three studies to exhibit significant differential expression between diseased and healthy samples. The analysis is implemented in the "R" programming language so any user can easily and freely reproduce the results.

Introduction

There are multiple RNASeq studies that have utilized expression profiling by high throughput sequencing to examine questions related to the expression effects of Covid-19. At present, three such studies contain a large number of samples that have led to numerous publications: GSE198449 with 1858(1–5), GSE215865 with 1392(6), and GSE212041 with 781(7) samples respectively. These studies represent an unprecedented public resource for Covid related research as well as an opportunity to conduct large scale biological analyses with a dataset that meets nearly any reasonable concern related to sample size.

With respect to public availability of data, the R programming language continues to democratize the world of sophisticated computational analyses as a "driver of reproducible science"(8) occupying a pivotal role in the space of complex biological analyses, empowering researchers across diverse backgrounds to delve into intricate biological datasets with precision, flexibility, and reproducibility(9,10). With a rich ecosystem of packages, particularly those made available through Bioconductor(10), and libraries tailored for statistical computing, data visualization, and machine learning, R provides an accessible platform for researchers to conduct analyses in genomics, bioinformatics, and other biological domains. The open-source nature of R fosters collaboration and knowledge-sharing within the scientific community, allowing users to leverage and contribute to a vast repository of tools and methods. R's adaptability and ease of use have made it a preferred choice for both wet-lab biologists and computational experts, enabling interdisciplinary collaboration and accelerating discoveries in areas such as genomics, metagenomics, and systems biology. As a result, R continues to even the playing field of biological research, making advanced analytical techniques more accessible and facilitating breakthroughs in our understanding of complex biological systems.

The synthesis of public RNASeq Covid-19 data with R-enabled analytical tooling is an obvious evolution of both public domain data-sharing initiatives and advanced computational capabilities. A large sample size makes statistical analyses of biological data a relatively simple affair. Combined with the analytical power of R, it has never been easier to conduct analyses that are consistent with best practices(11,12)

Intriguingly, to date no effort has been made to discover the level of agreement that exists among the data collected in the aforementioned RNASeq Covid studies. Here, a novel R-based analysis of data collected from these studies is presented, providing an example of how a relatively large scale analyses can easily be completed with modest computational resources (a laptop and a few afternoons of code writing).

Materials and methods.

All data considered in this study were identified from the Gene Expression Omnibus browser (https://www.ncbi.nlm.nih.gov/geo/browse) utilizing a simple search prompt, “Covid”, and by sorting the results with respect to “Series type(s)” - “Expression profiling by high throughput sequencing” was selected, and “Samples” – the studies with the three largest sample sizes were identified. This led to the selection of data from three studies: GSE198449, GSE215865, and GSE212041. Data were considered at the stage of annotation count data, TPM values where available (GSE212041) and raw counts where they were not (GSE198449 and GSE215865). All data were processed with R. Included with this submission are four workflow documents that will allow any user to completely reproduce the results for the individual analyses of the three RNASeq datasets as well as the final analysis that identified genes in common among all three datasets. Consistent with best practices, and to the degree possible, each dataset was processed with the same workflow that consisted of the following nine stages: 1. **Data**

and metadata download: Even at this initial step, analyses varied depending on the availability of metadata. In two cases, metadata were easily obtained from a single source; collection of metadata for the third study required gathering values from over 700 individual webpages. 2. **Harmonization of data and metadata**: Here data and metadata were matched to make sure that the data from each sample were matched with corresponding metadata. A small number of samples were culled from two studies based on apparent absence of metadata. In a third study more than two thirds of the samples were culled based on lack of correspondence between metadata and data. Harmonization of data with metadata required a custom approach to rectify sample names/IDs in each study. 3. **Examination of data distributions and summary statistics for the entire dataset and each individual sample**: Here samples were examined with an eye toward identification and removal of those that exhibited the characteristics of an outlier. While a number of outlier samples were identified in each study, just one sample in one study was removed. Samples were retained based on the notion that it would be possible to identify batch effects based on correlation between metadata and data in subsequent analysis steps. 4. **Preprocessing and attempted normalization**: Based on some of the most recent suggestions with respect to normalizing RNASeq based count data(12), an attempt was made to normalize all samples with a combination of quantile-based normalization and log-transformation. To reduce background noise, this procedure also eliminated extreme low counts (i.e. singletons). The code used to preprocess the data has several additional features that users can implement to fine tune their analyses; these are not discussed here 5. **Re-examination of data distributions and summary statistics for the entire dataset and each individual sample:** After pre-processing/normalization, the data were re-examined to determine if outlier samples were improved, and to determine the actual, not assumed, distribution of the data (essentially to determine if the preprocessed data exhibited a normal distribution or not). 6. **PCoA** Principal coordinate analysis was performed with one or more distance metrics on each of the datasets. Users can choose several additional distance metrics. The PCoAs were automatically rendered

and colored with respect to all collected metadata to enable identification of trends between the reduced dimension expression data and the metadata. The PCoAs were rendered both as static plots and as an interactive three dimensional plot. The interactive plot can be colored with respect to any metadata. 7. **Statistical analysis to identify the genes that exhibited the most significant differential expression between metadata identified disease and healthy states**: The same test was not used in each study. The test (Mann-Whitney or Kruskal-Wallis) was selected based on the following three criteria – normality (or lack thereof) of the data, number of groups, and the paired/unpaired nature of the samples. Users can perform additional available tests. The test was performed using metadata that indicated the Covid state (healthy or unhealthy) to select groupings of the samples.The exact details are provided in the workflow of each dataset. After initial statistical analysis, p-values were generated and corrected with Bonferroni and Benjamini-Hochberg methods to control for multiple testing. Additional p value adjustments are available. Data were then sorted and culled based on the Bonferroni adjusted p to produce a list of the ~5% of genes that exhibited the most significant differential expression between the diseased and healthy Covid states. Note that in one of the studies (GSE2120418) there were three Covid related states; in this case the Kruskal-Wallis test was used across all three groups. **Visualization of gene sets with heatmap-dendrogram and subsequent PCoA**: Heatmap-dendrograms were used to visualize the statistically culled expression data and corresponding metadata. In addition, a second round of PCoAs were calculated and visualized for the statistically culled genes identified in each dataset. In both cases visualizations used the same metadata used for the statistical analyses; users can generate PCoAs and heatmap-dendrograms colored for any of the available metadata. 9. **Annotation and preliminary pathway analysis**: Each study utilized ENSG Gene stable IDs to identify genes. These values were annotated with gene names and gene descriptions from the Ensembl biomart (https://useast.ensembl.org/biomart/martview). Annotated genes underwent a preliminary pathway analysis. **Final Analysis**. After this nine-step procedure was completed on each of the datasets an

additional analysis was conducted to identify the genes found in common among the statistically enriched sets for each study.

Results

In this analysis more than 4,000 initial samples were considered across three previously published studies. The data from each study was independently analyzed and the results of all three analyses were compared to identify 61 genes, 60 of which are annotated, that exhibit robust Covid-19 dependent differential expression (Supplemental Table 1). In one of the three analyses (GSE198449) segregation between diseased and healthy samples was obvious enough to be seen in the initial PCoA (Figure 1) and also in the statistically selected data presented in the heatmap-dendrograms (Figure 2). In the other two studies, separation between/among groups was less than obvious. Likewise, the dataset with obvious separation between diseased and healthy states exhibits a higher level of statistical fidelity (assessed as the distribution of Bonferroni corrected p values) than that observed in the other two (data not shown, but included in R-based analyses). The preliminary pathway analysis of the identified genes suggests that several pathways could be affected (Figure 3). Further investigation is in order.

Outlier samples are apparent in both the PCoAs (Figure 1 displays a static rendering of the PCoA for one dataset, the other two are included in R-based analyses) and heatmap-dendrograms (Figure 2). In large part these samples did not appear to correlate with any metadata related to the samples (R analyses include PCoAs colored by each class of metadata); some direct evidence of batch effects was observed (data not shown, but included in R analyses). However, it is clear that many samples exhibit characteristics that make them inconsistent with meaningful biological interpretation; these should be removed with objective, expert level criteria.

A warning to the reader, the code included in the R scripts is meant to be presented as a work in progress. It is facultative biologist code, not production level. The analysis required the development of much novel code. The majority was generalized, applicable to analysis of each dataset. However, some was customized for each analysis, to deal with idiosyncratic issues such as gathering of metadata from multiple sites or harmonization of data and metadata; these challenges required solutions that were particular to the analysis of each study. My hope is that the code itself (Supplemental Files GSE198449.R, GSE215865.R, GSE212041.R, and combine_covid.R along with the numerous packages and GitHub hosted code these workflows implement) will prove to be a valuable resource, a reference for conducting large scale re-analysis of biological count data with R.

Discussion

The Covid-19 based RNASeq studies utilized in the analyses presented here have been available to the public for some time. It is surprising that no previous published effort has been made to see if the data from these studies agree. My re-analysis of the data reveals a statistically robust (genes were ultimately filtered based on stringent Bonferroni corrected p values), but modestly sized (a total of 61 genes, 60 of which were successfully annotated) set of genes that should be of interest but that may suggest one or both of two non-exclusive possibilities. Either genes whose expression is genuinely affected by Covid are rare, or that Covid dependent signals in expression are subtle, damped out by other biological signals and noise in the data. The apparent discordance in the ability of each study to resolve Covid dependent expression may also suggest that differences in experimental protocol among the original studies led to divergence in their ability to observe Covid related genes. This is an intriguing possibility that deserves further study, well beyond the scope of this report.

The PCoAs and heatmap-dendrograms reveal that many samples in each study appear to be outliers,

samples with characteristics so extreme it is unlikely that any biologically meaningful conclusions can be drawn from them, even with best practices in data preprocessing/normalization. Simple summary statistic screening like that presented here, as well as much more sophisticated methods discussed elsewhere, should be utilized to screen out such samples. However, determination of which samples to keep should rely on expert advice, not arbitrary cutoffs, hence my reluctance to eliminate samples except in the case of missing metadata. Interestingly, a cursory review of the existing publications that utilized the datasets considered here revealed no such screening procedures to exclude outlier samples.

While data and metadata are publicly available, there are few apparent rules that dictate exactly what type of data is available. Indeed, the variety of data (TPM in one study vs raw counts in the other two), and particularly the amount of effort required to match metadata with data was somewhat surprising. In one study, more than 1000 samples were not used due to apparent lack or disagreement of metadata with respect to sample names. This may very well have been due to over-site on my part (please revue the work and by all means let me know – use the GitHub link below), but more uniform standards for reporting sample names and corresponding metadata would certainly enhance efforts to reuse data like that reported here.

The analyses here were not utilized to compare the different methods used to generate the count data. There could be obvious value in such a comparison, but I leave that to other, more capable researchers.

Here a unique set of genes that exhibit Covid-19 dependent differential expression from previously published data is presented in hopes that this robust gene set will prove useful to the scientific and medical communities. As a potential template for further more improved and more sophisticated studies, I provide my complete analysis in an easily reproduced format (R scripts), making replication,

and more importantly improvement, of this work a trivial matter.

Conclusion

The included workflow documents will allow any user to easily reproduce the entirety of the analyses presented here - including all steps, intermediary and final data products, visualizations etc. not directly included in this article. For those interested, I recommend using RStudio or a similar integrated development environment (IDE) to process through the workflow documents. In this way, the entire analysis can be completed in minutes. You will also be able to observe the detailed statistical results for each study as well as large scale and interactive visualizations that could not be included on these pages. Unfortunately, a few analysis steps cannot be completed directly in R; in these instances the workflow documents provide detailed instructions. For the sake of convenience, the code used in this study is available as supplemental material but also as a repository on GitHub (https://github.com/DrOppenheimer/covid_play_R).

In addition to the code used to analyze the data, I present a list of 74 genes found to exhibit highly significant Covid-19 dependent differential expression across three RNASeq based studies. My hope is this list will be of use to the scientific and medical communities.

Acknowledgments

I would like to express my heartfelt gratitude to my wife, Dr. Jennifer Kossoris, for her unwavering support throughout the preparation of this manuscript. Her invaluable comments and insights greatly enhanced the quality of the work. I am deeply appreciative of her patience, understanding, and encouragement during the entire process. This work would not have been possible without her steadfast support.

References

1. Mao W, Miller CM, Nair VD, Ge Y, Amper MAS, Cappuccio A, et al. A methylation clock model of mild SARS-CoV-2 infection provides insight into immune dysregulation. Mol Syst Biol. 2023;19: e11361. doi:10.15252/msb.202211361

2. Soares-Schanoski A, Sauerwald N, Goforth CW, Periasamy S, Weir DL, Lizewski S, et al. Asymptomatic SARS-CoV-2 Infection Is Associated With Higher Levels of Serum IL-17C, Matrix Metalloproteinase 10 and Fibroblast Growth Factors Than Mild Symptomatic COVID-19. Front Immunol. 2022;13: 821730. doi:10.3389/fimmu.2022.821730

3. Zhang Z, Sauerwald N, Cappuccio A, Ramos I, Nair VD, Nudelman G, et al. Blood RNA alternative splicing events as diagnostic biomarkers for infectious disease. Cell Rep Methods. 2023;3: 100395. doi:10.1016/j.crmeth.2023.100395

4. Ren J, Zhang Y, Guo W, Feng K, Yuan Y, Huang T, et al. Identification of Genes Associated with the Impairment of Olfactory and Gustatory Functions in COVID-19 via Machine-Learning Methods. Life (Basel). 2023;13: 798. doi:10.3390/life13030798

5. Sauerwald N, Zhang Z, Ramos I, Nair VD, Soares-Schanoski A, Ge Y, et al. Pre-infection antiviral innate immunity contributes to sex differences in SARS-CoV-2 infection. Cell Syst. 2022;13: 924-931.e4. doi:10.1016/j.cels.2022.10.005

6. Thompson RC, Simons NW, Wilkins L, Cheng E, Del Valle DM, Hoffman GE, et al. Molecular states during acute COVID-19 reveal distinct etiologies of long-term sequelae. Nat Med. 2023;29: 236–246. doi:10.1038/s41591-022-02107-4

7. LaSalle TJ, Gonye ALK, Freeman SS, Kaplonek P, Gushterova I, Kays KR, et al. Longitudinal characterization of circulating neutrophils uncovers phenotypes associated with severity in hospitalized COVID-19 patients. Cell Rep Med. 2022;3: 100779. doi:10.1016/j.xcrm.2022.100779

8. Giorgi FM, Ceraolo C, Mercatelli D. The R Language: An Engine for Bioinformatics and Data

Science. Life (Basel). 2022;12: 648. doi:10.3390/life12050648

9. Jalal H, Pechlivanoglou P, Krijkamp E, Alarid-Escudero F, Enns E, Hunink MGM. An Overview of R in Health Decision Sciences. Med Decis Making. 2017;37: 735–746. doi:10.1177/0272989X16686559

10. Sepulveda JL. Using R and Bioconductor in Clinical Genomics and Transcriptomics. J Mol Diagn. 2020;22: 3–20. doi:10.1016/j.jmoldx.2019.08.006

11. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17: 13. doi:10.1186/s13059-016-0881-8

12. Abrams ZB, Johnson TS, Huang K, Payne PRO, Coombes K. A protocol to evaluate RNA sequencing normalization methods. BMC Bioinformatics. 2019;20: 679. doi:10.1186/s12859-019-3247-x

13. Kolberg L, Raudvere U, Kuzmin I, Vilo J, Peterson H. gprofiler2– an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler. F1000Research. 2020;9 (ELIXIR).

242

243 **Supporting Information**

244 **Supplementary_Table_1**. Genes that exhibit significant differential expression across all three
245 analyzed datasets

246 **GSE198449.R** Complete analysis of dataset GSE198449

247 **GSE215865.R** Complete analysis of dataset GSE215865

248 **GSE212041.R** Complete analysis of dataset GSE212041

249 **combine_covid.R** Analysis identifying genes in common among all three fully processed datasets.

250

251

Figure and Tables

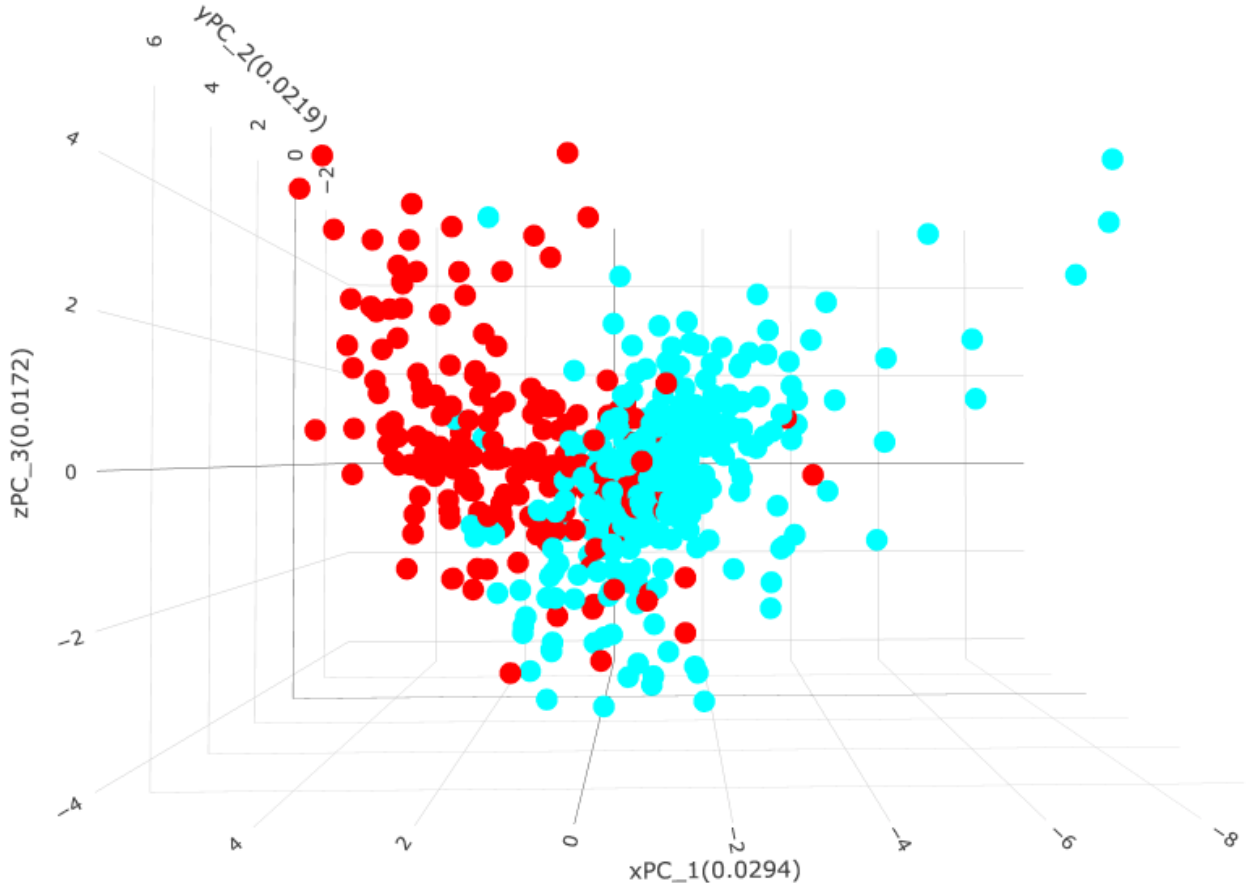
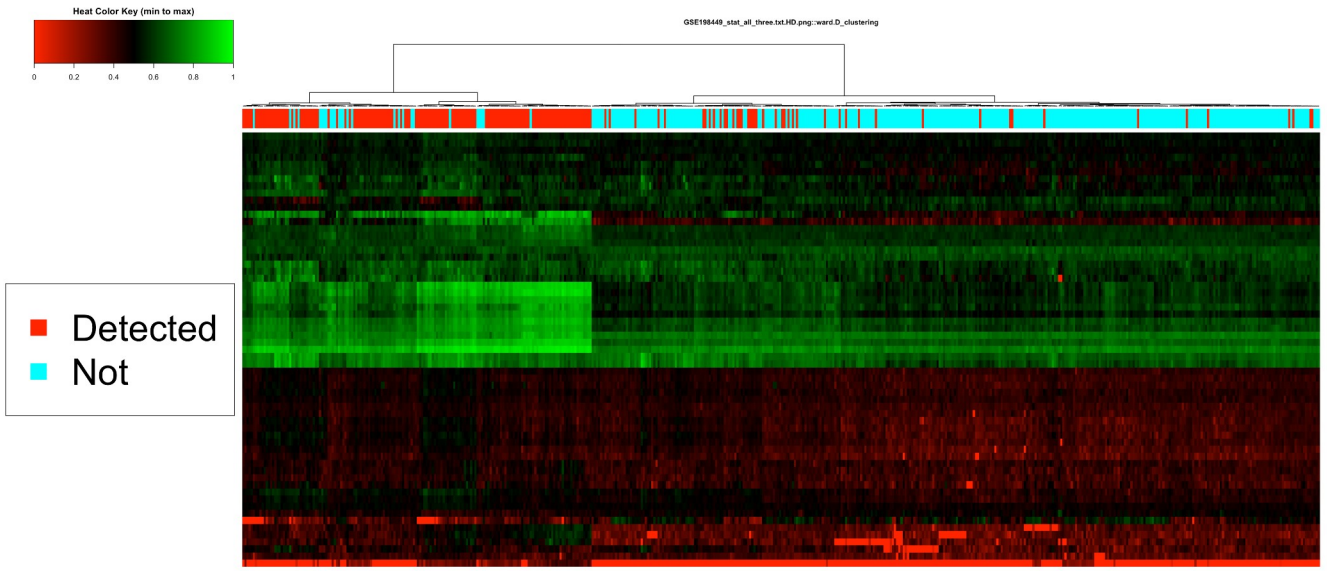
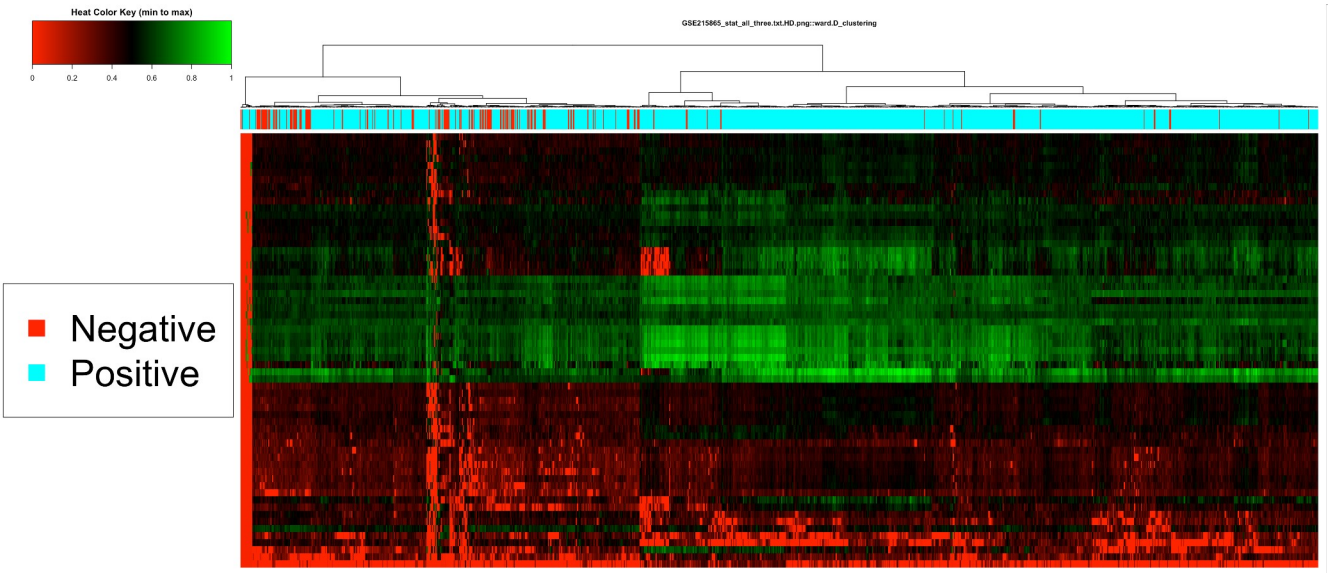


Figure 1. PCoA of "PCR test for SARS-Cov-2" , "Not" vs "Detected" from GSE198449

A Euclidean distance-based PCoA was calculated from normalized count values as described above. Metadata from the original study was used to group and color samples with respect to PCR defined Covid disease state. Three dimensional PCoAs based on the first three eigen vectors were visualized as static images (automatically colored by all metadata) or as an interactive 3d PCoA generated from selected metadata. Coordinate values indicate the scaled eigen values for each coordinate; these can be interpreted as % variation displayed, i.e. - xPC_1 displays 3% of all detected variation. Here there is clear, but not perfect separation between Covid samples labeled as "Not" and "Detected".


Heat Color Key (min to max)
0
0.2
0.4
0.6
0.8
1
GSE198449_stat_all_three.txt.HD.png::ward.D_clustering
Detected
Not


Heat Color Key (min to max)
0
0.2
0.4
0.6
0.8
1
GSE215865_stat_all_three.txt.HD.png::ward.D_clustering
Negative
Positive

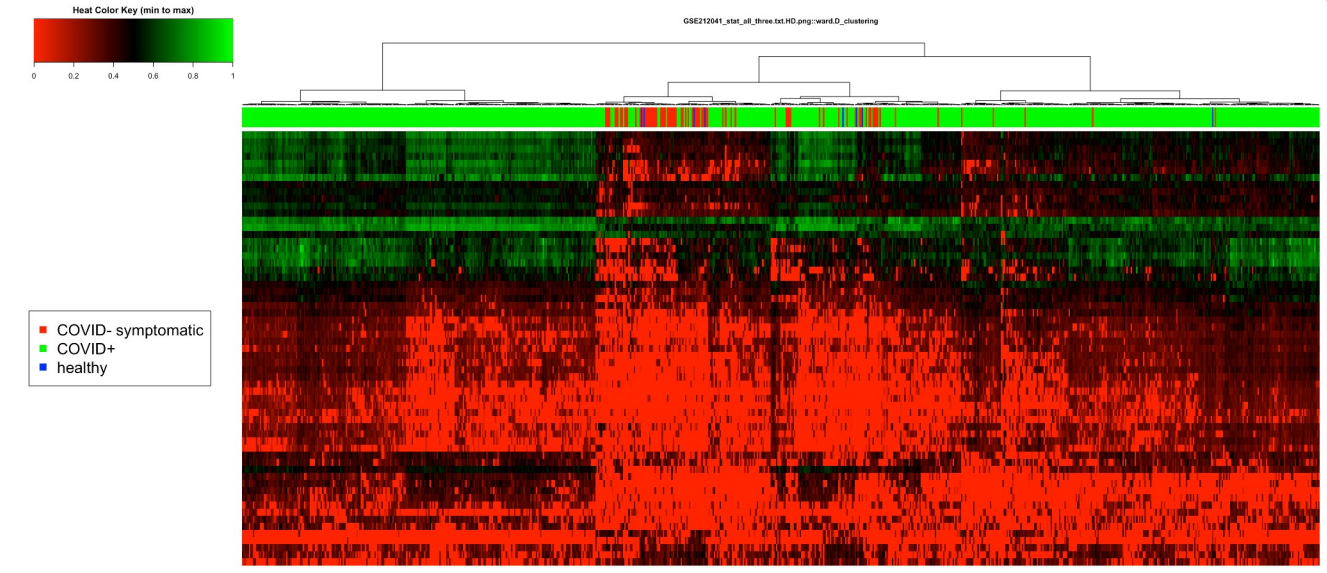
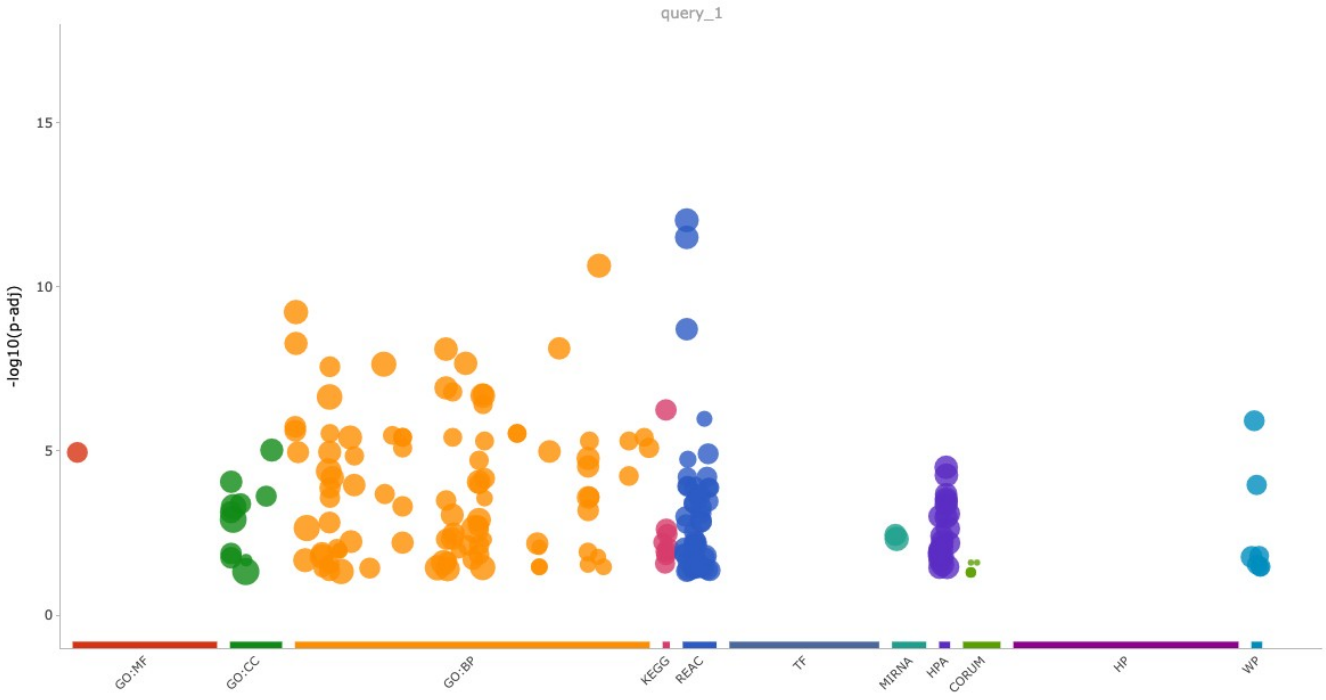

Heat Color Key (min to max)
0
0.2
0.4
0.6
0.8
1
GSE212041_stat_all_three.txt.HD.png::ward.D_clustering
COVID- symptomatic
COVID+
healthy

Figure 2. Heatmap-dendrogrms for 74 genes found to exhibit significant differential expression across all three datasets.

Heatmap-dendrograms display the normalized (as discussed above) expression for 74 genes that exhibit statistically significant Covid-19 dependent differential expression across all three studies. The expression observed in each individual study is shown. Top GSE198449(506 samples), middle GSE215865(1391 samples), and bottom GSE212041(781 samples). The color bar at the top of each heatmap-dendrogram as well as the legend at the left display the selected metadata for each sample.



284 **Figure 3. Pathway Analysis Visualization of the 74 Discovered Genes**

285 ENSG based annotations were used to produce a pathway analysis utilizing the 74 discovered genes
286 and the gprofiler2 package for R(13). A Manhattan plot of enrichment results from gprofler2. An
287 interactive version of this visualization is available through the R analysis.