

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

3

4 Kevin P. Keegan, Ph.D.^{1*}

5

6 ¹Independent researcher, Glen Ellyn, IL, USA

7 * Corresponding Author

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

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25 **Abstract**

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

43

44

45

46

47

48

49 **Introduction**

50 There are multiple RNASeq studies that have utilized expression profiling by high throughput
51 sequencing to examine questions related to the expression effects of Covid-19. At present, three such
52 studies contain a large number of samples that have led to numerous publications: GSE198449 with
53 1858(1–5), GSE215865 with 1392(6), and GSE212041 with 781(7) samples respectively. These
54 studies represent an unprecedented public resource for Covid related research as well as an opportunity
55 to conduct large scale biological analyses with a dataset that meets nearly any reasonable concern
56 related to sample size.

57

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

72

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

78
79 Intriguingly, to date no effort has been made to discover the level of agreement that exists among the
80 data collected in the aforementioned RNASeq Covid studies. Here, a novel R-based analysis of data
81 collected from these studies is presented, providing an example of how a relatively large scale analyses
82 can easily be completed with modest computational resources (a laptop and a few afternoons of code
83 writing).

84

85 **Materials and methods.**

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

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

121 and colored with respect to all collected metadata to enable identification of trends between the reduced
122 dimension expression data and the metadata. The PCoAs were rendered both as static plots and as an
123 interactive three dimensional plot. The interactive plot can be colored with respect to any metadata. 7.

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

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

147

148 **Results**

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

160

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

168

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

178

179 **Discussion**

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

191

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

193 samples with characteristics so extreme it is unlikely that any biologically meaningful conclusions can
194 be drawn from them, even with best practices in data preprocessing/normalization. Simple summary
195 statistic screening like that presented here, as well as much more sophisticated methods discussed
196 elsewhere, should be utilized to screen out such samples. However, determination of which samples to
197 keep should rely on expert advice, not arbitrary cutoffs, hence my reluctance to eliminate samples
198 except in the case of missing metadata. Interestingly, a cursory review of the existing publications that
199 utilized the datasets considered here revealed no such screening procedures to exclude outlier samples.
200

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

209
210 The analyses here were not utilized to compare the different methods used to generate the count data.
211 There could be obvious value in such a comparison, but I leave that to other, more capable researchers.
212

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

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

218

219 **Conclusion**

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

230

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

234

235 **Declarations**

236 **Ethics approval and consent to participate**

237 Not applicable.

238 **Consent for publication**

239 Not applicable.

240 **Availability of data and materials**

241 All data were obtained from public sources as outlined in the manuscript. The complete analyses are
242 publicly available on this GitHub repository: https://github.com/DrOppenheimer/covid_play_R.

243 **Competing interests**

244 Not applicable.

245 **Funding**

246 Not applicable.

247 **Authors' contributions**

248 K.P.K. was responsible for the entirety of this work. He conducted all research, data analysis, and
249 manuscript preparation.

250 **Acknowledgements**

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

256

257

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).

259

260 **Supporting Information**

261 **Supplementary_Table_1.** Genes that exhibit significant differential expression across all three
262 analyzed datasets

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

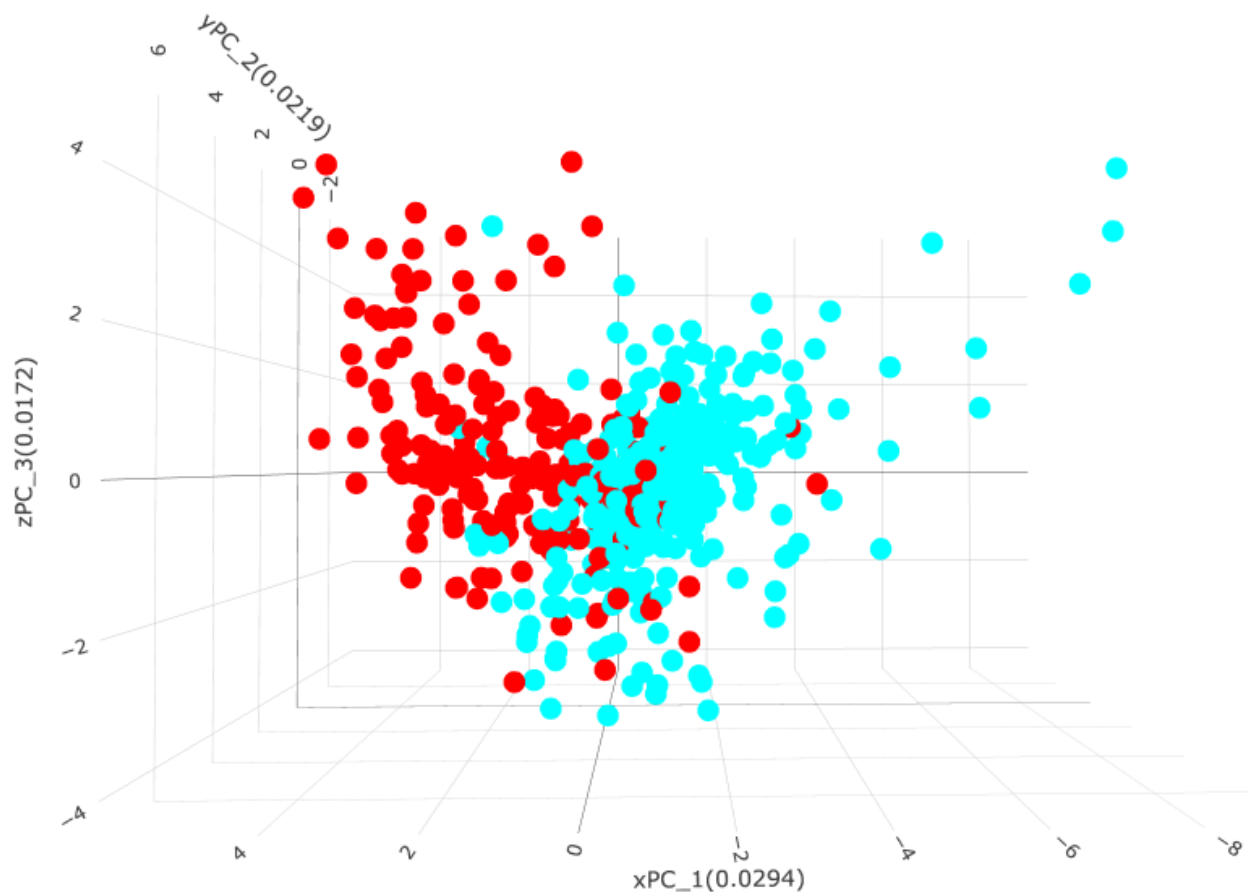
264 **GSE215865.R** Complete analysis of dataset GSE215865

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

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

267

268



270 **Figure 1. PCoA of “PCR test for SARS-Cov-2” , “Not” vs “Detected” from GSE198449**

271 A Euclidean distance-based PCoA was calculated from normalized count values as described above.

272 Metadata from the original study was used to group and color samples with respect to PCR defined

273 Covid disease state. Three dimensional PCoAs based on the first three eigen vectors were visualized as

274 static images (automatically colored by all metadata) or as an interactive 3d PCoA generated from

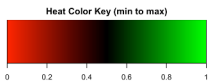
275 selected metadata. Coordinate values indicate the scaled eigen values for each coordinate; these can be

276 interpreted as % variation displayed, i.e. - xPC_1 displays 3% of all detected variation. Here there is

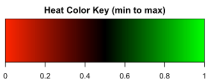
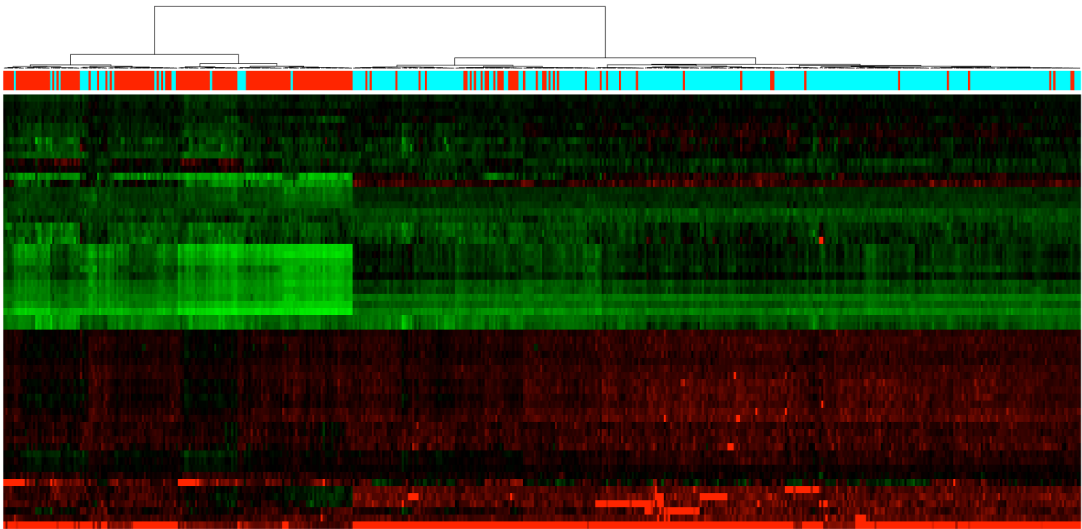
277 clear, but not perfect separation between Covid samples labeled as “Not” and “Detected”.

278

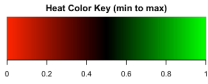
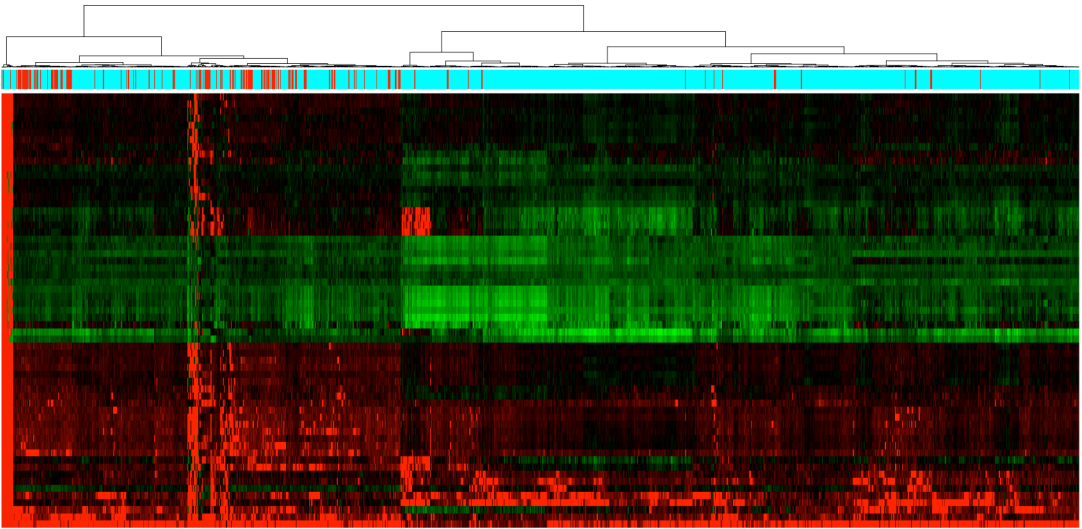
279



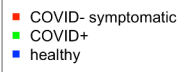
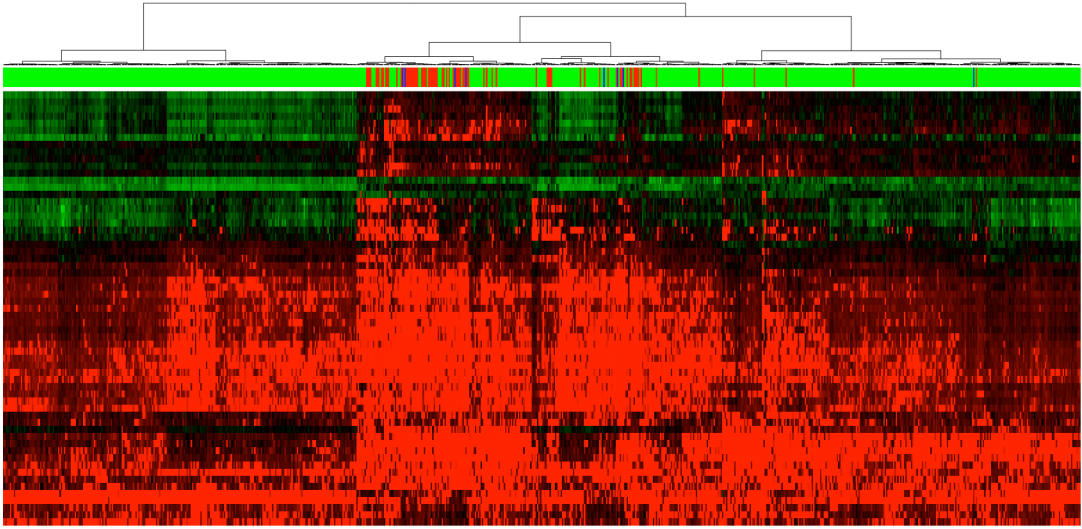
GSE118448_stat_all_three.txt.HD.png:ward.D_clustering



GSE215865_stat_all_three.txt.HD.png:ward.D_clustering



GSE212041_stat_all_three.txt.HD.png:ward.D_clustering



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

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

288

289

290

291

292

293

294

295

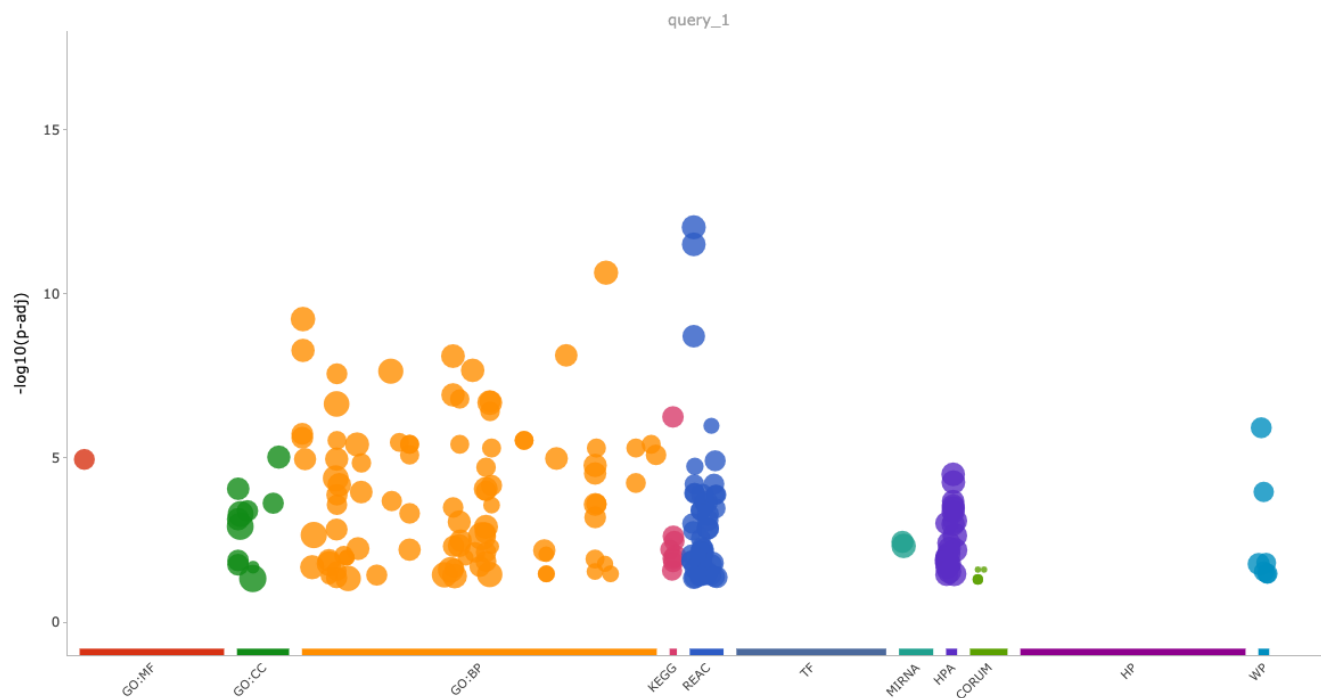
296

297

298

299

300



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

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