Ensemble Methods for Outlier Detection (EnsMOD) User Guide

Nathan P. Manes, Jian Song, Aleksandra Nita-Lazar

Laboratory of Immune System Biology, National Institute of Allergy
and Infectious Diseases, National Institutes of Health, Bethesda, MD, USA

Contact us at EnsMOD-team@nih.gov

Introduction

Detection of omics sample outliers is important for preventing erroneous biological conclusions, developing robust experimental protocols, and discovering rare biological states. Two recent publications describe robust algorithms for detecting transcriptomic sample outliers (Chen et al 2020; Selicato et al 2021).

For the first algorithm, PcaGrid was used (Chen et al 2020). For the other algorithm, hierarchical cluster analysis (HCA) and ROBPCA were integrated (Selicato et al 2021). Unfortunately, neither of these two algorithms had been incorporated into a software program accessible to omics scientists without a strong background in bioinformatics or biostatistics.

Ensemble Methods for Outlier Detection (EnsMOD) incorporates both algorithms. EnsMOD plots density curves of each sample to visualize anomalies, calculates how closely the quantitation variation follows a normal distribution, performs hierarchical cluster analyses to calculate how closely the samples cluster with each other, and performs robust principal component analyses to statistically test if any sample is an outlier. EnsMOD is open-source and freely available (https://github.com/niaid/EnsMOD).

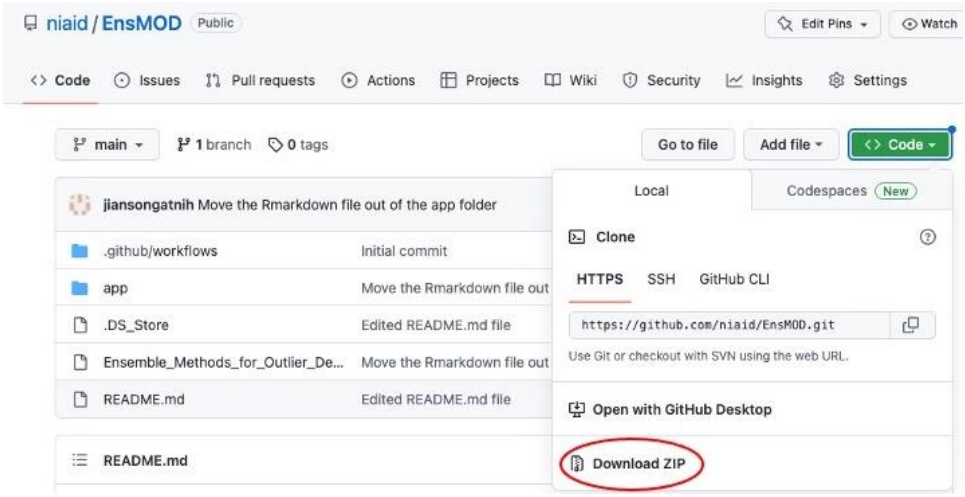
Installing and Running EnsMOD

EnsMOD is provided both as a stand-alone Rmarkdown and as an application with a graphical user interface. Both versions have exactly the same functionality. In order to run EnsMOD, the following steps are needed to set up EnsMOD on your computer:

1. Install R (https://www.r-project.org/).
2. Install RStudio (https://www.rstudio.com/).
3. **Acquire EnsMOD.** EnsMOD is freely available at https://github.com/niaid/EnsMOD.
    a. To use the Rmarkdown version, simply download "*Ensemble_Methods_for_Outlier_Detection_v2_0_stand_alone.Rmd*" and open it using RStudio.

b. To use the application version, download it as a ZIP file (click the green "Code" button, and then click "Download ZIP"). Decompress the ZIP file into a directory, and open the */app/app.R* file in RStudio.

c.



4. **Acquire the required EnsMOD R Packages.** The script and application versions automatically install and update all of their required R packages.

    a. Alternatively, to manually install the R packages required by the script version, open RStudio and select "Tools" and "Install Packages..." from the drop-down menu, and install the required packages (including dependencies): *BiocManager*, *cluster*, *factoextra*, *fitdistrplus*, *ggraph*, *gplots*, *RColorBrewer*, *rospca*, *rrcov*, and *tidyverse*; use the Console to run "*BiocManager::install("limma")*" to install the *limma* package.

    b. It is possible that antivirus software will need to be paused during these steps.

    c. To confirm that the "*cluster*" package was successfully installed, try to load it by running "*library(cluster)*" using the Console, and similarly for the other packages.

    d. In addition to the above R packages, the application version also requires: *shiny*, *shinyjs*, *xfun*, *DT*, *readr*, *dplyr*, *data.table*, *reshape2*, *htmltools*, and *readxl*.

5. **Provide a table of input data.**

    a. The input table needs to be in an XLSX file.

    b. For the application version, the input XLSX file is opened using the GUI (described below).

    c. For the Rmarkdown version, the input XLSX file needs to be named "*Gene_Expression_Table.xlsx*", and it needs to be located in the same directory as the EnsMOD Rmarkdown file.

    d. The XLSX file should contain only one worksheet and only one table.

i. In this table of gene/protein/etc. abundance data, each column corresponds to a sample, and each row corresponds to a gene (or a protein, or a metabolite, etc., depending on the omics dataset type).

ii. The first row (the header) should contain unique identifiers of the samples.

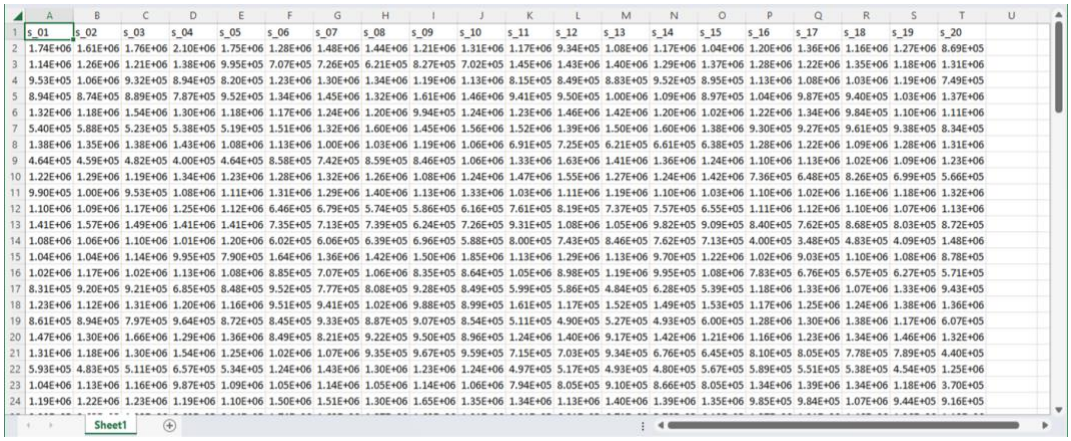
iii. The first column is not for gene/protein/etc. unique identifiers (they are not needed and should be removed). All of the columns correspond to samples and contain abundance values.

e. Rows may contain missing values (such as "NaN"), but these rows will be excluded from the analyses (note that "0" is not treated as a missing value).

f. Four example datasets are provided at https://github.com/niaid/EnsMOD/tree/main/app/EnsMOD_Examples. The table below is of the simulated proteomics dataset (described below).

g. Data imputation (before using EnsMOD) might be necessary if most of the rows contain one or more missing values. We recommend against using EnsMOD to analyze sparse datasets with missing measured (i.e., non-imputed) values >50%, and we caution that data imputation might negatively affect omics outlier detection in general.

h. The overall quantitation variation is assumed to be normally distributed (discussed in the "Data Normality" subsection below). If this requires a transformation of the data (e.g., log-transformation), it must be performed separately prior to the EnsMOD analysis.



s_01	s_02	s_03	s_04	s_05	s_06	s_07	s_08	s_09	s_10	s_11	s_12	s_13	s_14	s_15	s_16	s_17	s_18	s_19	s_20
1.74E+06	1.61E+06	1.76E+06	2.10E+06	1.75E+06	1.28E+06	1.48E+06	1.44E+06	1.21E+06	1.31E+06	1.17E+06	9.34E+05	1.08E+06	1.17E+06	1.04E+06	1.20E+06	1.36E+06	1.16E+06	1.27E+06	8.69E+05
1.14E+06	1.26E+06	1.21E+06	1.38E+06	9.95E+05	7.07E+05	7.26E+05	6.21E+05	8.27E+05	7.02E+05	1.45E+06	1.43E+06	1.40E+06	1.29E+06	1.37E+06	1.28E+06	1.22E+06	1.35E+06	1.18E+06	1.31E+06
9.53E+05	1.06E+06	9.32E+05	8.94E+05	8.20E+05	1.23E+06	1.30E+06	1.34E+06	1.19E+06	1.13E+06	8.15E+05	8.49E+05	8.83E+05	9.52E+05	8.95E+05	1.13E+06	1.08E+06	1.03E+06	1.19E+06	7.49E+05
8.94E+05	8.74E+05	8.89E+05	7.87E+05	9.52E+05	1.34E+06	1.45E+06	1.32E+06	1.61E+06	1.46E+06	9.41E+05	9.50E+05	1.00E+06	1.09E+06	8.97E+05	1.04E+06	9.87E+05	9.40E+05	1.03E+06	1.37E+06
1.32E+06	1.18E+06	1.54E+06	1.30E+06	1.18E+06	1.17E+06	1.24E+06	1.20E+06	9.94E+05	1.24E+06	1.23E+06	1.46E+06	1.42E+06	1.20E+06	1.02E+06	1.22E+06	1.34E+06	9.84E+05	1.10E+06	1.11E+06
5.40E+05	5.88E+05	5.23E+05	5.38E+05	5.19E+05	1.51E+06	1.32E+06	1.60E+06	1.45E+06	1.56E+06	1.52E+06	1.39E+06	1.50E+06	1.60E+06	1.38E+06	9.30E+05	9.27E+05	9.61E+05	9.38E+05	8.34E+05
1.38E+06	1.35E+06	1.38E+06	1.43E+06	1.08E+06	1.13E+06	1.00E+06	1.03E+06	1.19E+06	1.06E+06	6.91E+05	7.25E+05	6.21E+05	6.61E+05	6.38E+05	1.28E+06	1.22E+06	1.09E+06	1.28E+06	1.31E+06
4.64E+05	4.59E+05	4.82E+05	4.00E+05	4.64E+05	8.58E+05	7.42E+05	8.59E+05	8.46E+05	1.06E+06	1.33E+06	1.63E+06	1.41E+06	1.36E+06	1.24E+06	1.10E+06	1.13E+06	1.02E+06	1.09E+06	1.23E+06
1.22E+06	1.29E+06	1.19E+06	1.34E+06	1.23E+06	1.28E+06	1.32E+06	1.26E+06	1.08E+06	1.24E+06	1.47E+06	1.55E+06	1.27E+06	1.24E+06	1.42E+06	7.36E+05	6.48E+05	8.26E+05	6.99E+05	5.66E+05
9.90E+05	1.00E+06	9.53E+05	1.08E+06	1.11E+06	1.31E+06	1.29E+06	1.40E+06	1.13E+06	1.33E+06	1.03E+06	1.11E+06	1.19E+06	1.10E+06	1.03E+06	1.10E+06	1.02E+06	1.16E+06	1.18E+06	1.32E+06
1.10E+06	1.09E+06	1.17E+06	1.25E+06	1.12E+06	6.46E+05	6.79E+05	5.74E+05	5.86E+05	6.16E+05	7.61E+05	8.19E+05	7.37E+05	7.57E+05	6.55E+05	1.11E+06	1.12E+06	1.10E+06	1.07E+06	1.13E+06
1.41E+06	1.57E+06	1.49E+06	1.41E+06	1.41E+06	7.35E+05	7.13E+05	7.39E+05	6.24E+05	7.26E+05	9.31E+05	1.08E+06	1.05E+06	9.82E+05	9.09E+05	8.40E+05	7.62E+05	8.68E+05	8.03E+05	8.72E+05
1.08E+06	1.06E+06	1.10E+06	1.01E+06	1.20E+06	6.02E+05	6.06E+05	6.39E+05	6.96E+05	5.88E+05	8.00E+05	7.43E+05	8.46E+05	7.62E+05	7.13E+05	4.00E+05	3.48E+05	4.83E+05	4.09E+05	1.48E+06
1.04E+06	1.04E+06	1.14E+06	9.95E+05	7.90E+05	1.64E+06	1.36E+06	1.42E+06	1.50E+06	1.85E+06	1.13E+06	1.29E+06	1.13E+06	9.70E+05	1.22E+06	1.02E+06	9.03E+05	1.10E+06	1.08E+06	8.78E+05
1.02E+06	1.17E+06	1.02E+06	1.13E+06	1.08E+06	8.85E+05	7.07E+05	1.06E+06	8.35E+05	8.64E+05	1.05E+06	8.98E+05	1.19E+06	9.95E+05	1.08E+06	7.83E+05	6.76E+05	6.57E+05	6.27E+05	5.71E+05
8.31E+05	9.20E+05	9.21E+05	6.85E+05	8.48E+05	9.52E+05	7.77E+05	8.08E+05	9.28E+05	8.49E+05	5.99E+05	5.86E+05	4.84E+05	6.28E+05	5.39E+05	1.18E+06	1.33E+06	1.07E+06	1.33E+06	9.43E+05
1.23E+06	1.12E+06	1.31E+06	1.20E+06	1.16E+06	9.51E+05	9.41E+05	1.02E+06	9.88E+05	8.99E+05	1.61E+05	1.17E+05	1.52E+05	1.49E+05	1.53E+05	1.17E+06	1.25E+06	1.24E+06	1.38E+06	1.36E+06
8.61E+05	8.94E+05	7.97E+05	9.64E+05	8.72E+05	8.45E+05	9.33E+05	8.87E+05	9.07E+05	8.54E+05	5.11E+05	4.90E+05	5.27E+05	4.93E+05	6.00E+05	1.28E+06	1.30E+06	1.38E+06	1.17E+06	6.07E+05
1.47E+06	1.30E+06	1.66E+06	1.29E+06	1.36E+06	8.49E+05	8.21E+05	9.22E+05	9.50E+05	8.96E+05	1.24E+06	1.40E+06	9.17E+05	1.42E+06	1.21E+06	1.16E+06	1.23E+06	1.34E+06	1.46E+06	1.32E+06
1.31E+06	1.18E+06	1.30E+06	1.54E+06	1.25E+06	1.02E+06	1.07E+06	9.35E+05	9.67E+05	9.59E+05	7.15E+05	7.03E+05	9.34E+05	6.76E+05	6.45E+05	8.10E+05	8.05E+05	7.78E+05	7.89E+05	4.40E+05
5.93E+05	4.83E+05	5.11E+05	6.57E+05	5.34E+05	1.24E+06	1.43E+06	1.30E+06	1.23E+06	1.24E+06	4.97E+05	5.17E+05	4.93E+05	4.80E+05	5.67E+05	5.89E+05	5.51E+05	5.38E+05	4.54E+05	1.25E+06
1.04E+06	1.13E+06	1.16E+06	9.87E+05	1.09E+06	1.05E+06	1.14E+06	1.05E+06	1.14E+06	1.06E+06	7.94E+05	8.05E+05	9.10E+05	8.66E+05	8.05E+05	1.34E+06	1.39E+06	1.34E+06	1.18E+06	3.70E+05
1.19E+06	1.22E+06	1.23E+06	1.19E+06	1.10E+06	1.50E+06	1.51E+06	1.30E+06	1.65E+06	1.35E+06	1.34E+06	1.13E+06	1.40E+06	1.39E+06	1.35E+06	9.85E+05	9.84E+05	1.07E+06	9.44E+05	9.16E+05

6. **Adjust the four statistical parameters (optional).**

a. The four outlier detection stringency parameters are:

i. The minimum CCC threshold

ii. The maximum SC threshold

iii. The robpca probabilistic threshold

iv. The PcaGrid probabilistic threshold

b. These four threshold values are described in the "Interpretation of the EnsMOD Output" section below.

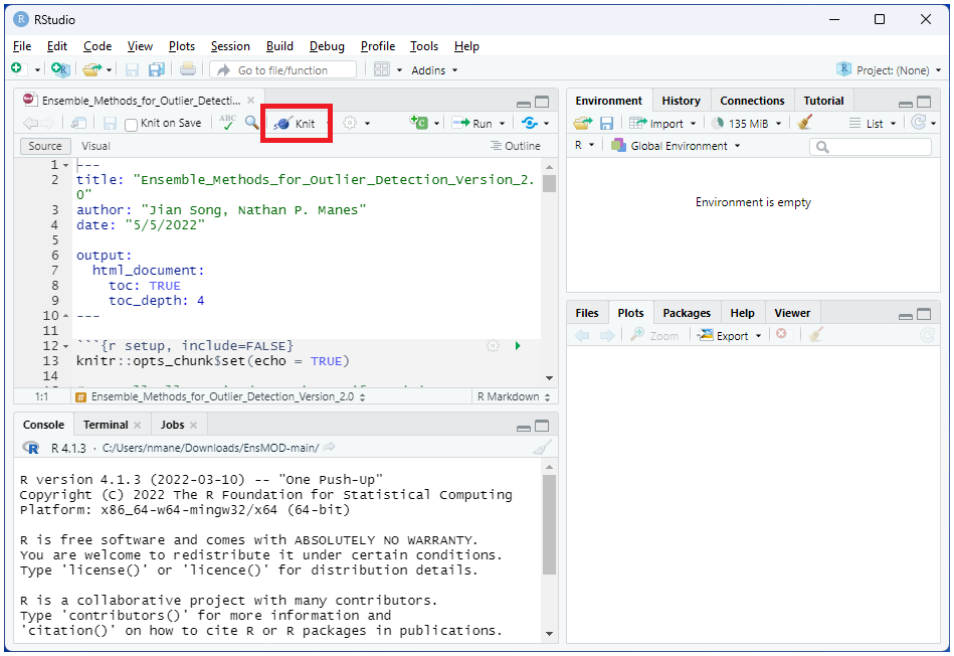
c. These four threshold values can be loosened or tightened by the user in RStudio (Rmarkdown version; these values are defined at the beginning of the Rmarkdown file) or using the GUI (application version).

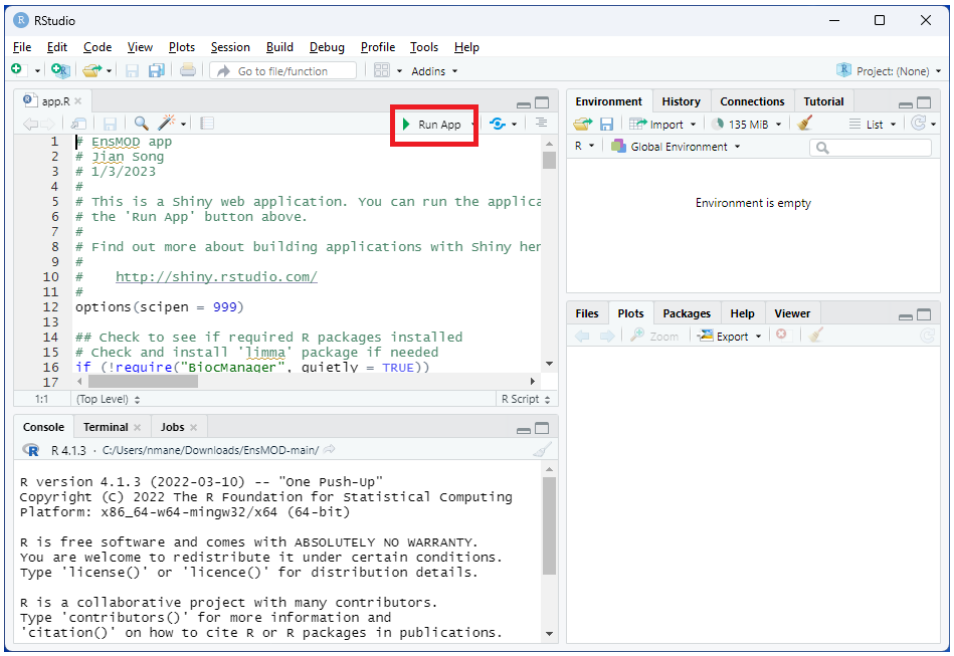
7. **Run EnsMOD.**

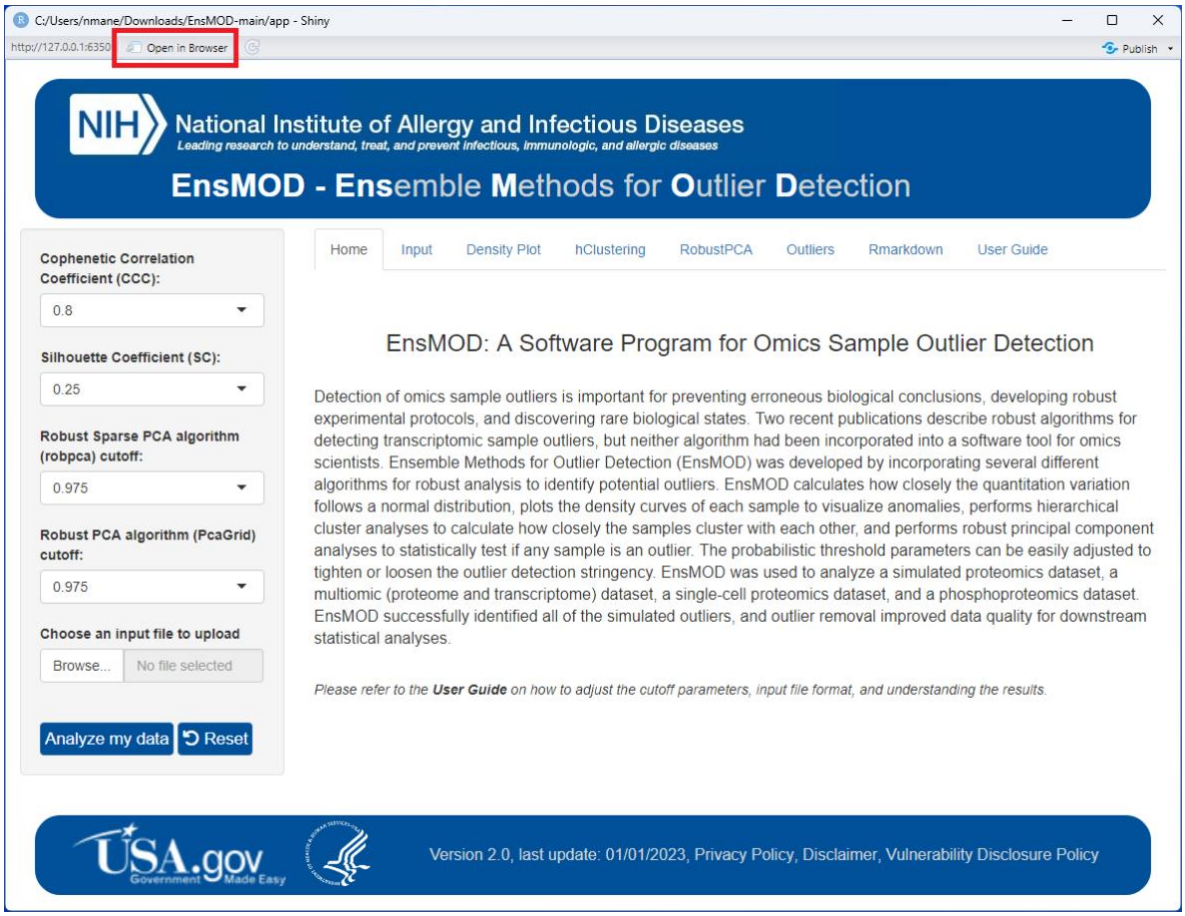
   a. To run the stand-alone Rmarkdown version, open the '*Ensemble_Methods_for_Outlier_Detection_v2_0_stand_alone.Rmd*' file in RStudio and click the "Knit" button.

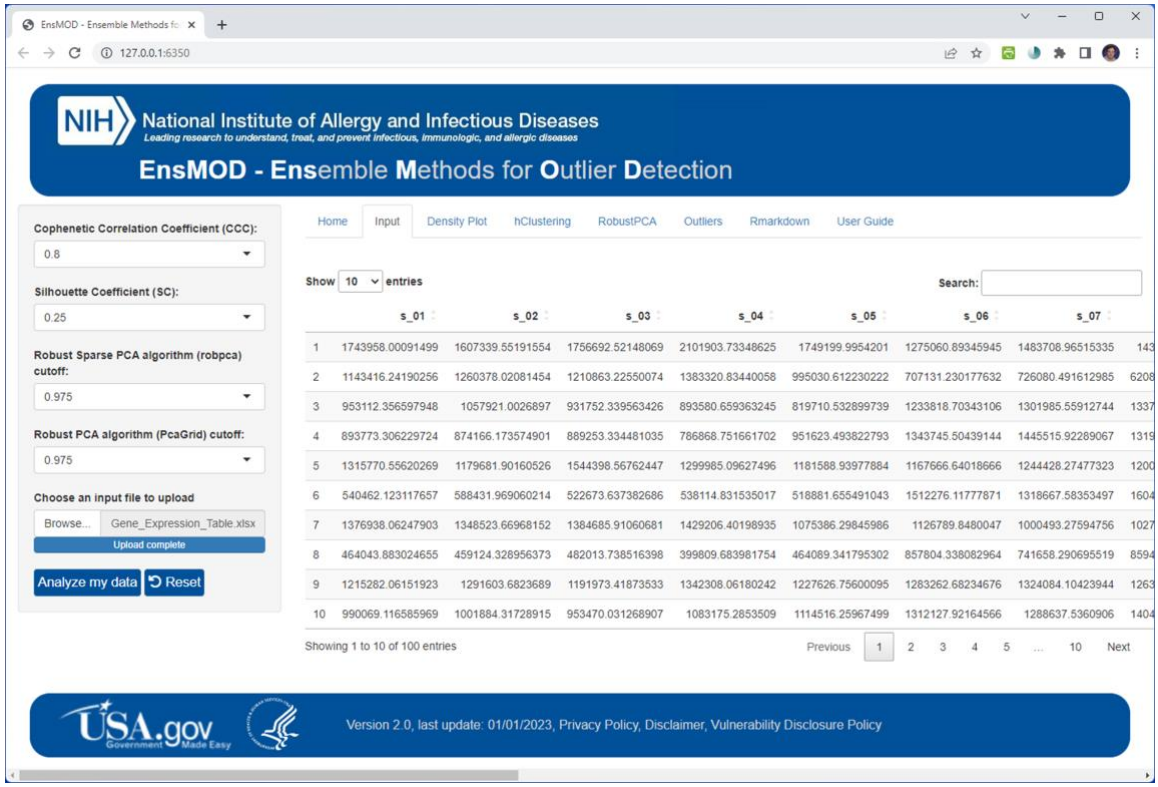
   b. To run EnsMOD as a Shiny application, open the '*app.R*' in the app/ folder in RStudio, click "Run App". It opens in the RStudio browser. Optional: click "Open in Browser" to open EnsMOD in the default www browser.

   c. For the example datasets (https://github.com/niaid/EnsMOD/tree/main/app/EnsMOD_Examples), the analysis runtimes ranged from approximately one minute to five hours (the SCoPE2 dataset took five hours; the input table dimensions were 1,490 samples by 731 proteins).




RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
app.R
Run App
# EnsMOD app
# Jian Song
# 1/3/2023
# This is a Shiny web application. You can run the applica
# the 'Run App' button above.
# Find out more about building applications with Shiny her
# http://shiny.rstudio.com/
options(scipen = 999)
Check to see if required R packages installed
# Check and install 'limma' package if needed
if (!require("BiocManager", quietly = TRUE))
Environment History Connections Tutorial
Environment is empty
Files Plots Packages Help Viewer
Console Terminal Jobs
R 4.1.3 · C:/Users/nmane/Downloads/EnsMOD-main/
R version 4.1.3 (2022-03-10) -- "One Push-Up"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.


C:/Users/nmane/Downloads/EnsMOD-main/app - Shiny
http://127.0.0.1:6350
Open in Browser
Publish
NIH National Institute of Allergy and Infectious Diseases
Leading research to understand, treat, and prevent infectious, immunologic, and allergic diseases
EnsMOD - Ensemble Methods for Outlier Detection
Cophenetic Correlation Coefficient (CCC):
0.8
Silhouette Coefficient (SC):
0.25
Robust Sparse PCA algorithm (robpca) cutoff:
0.975
Robust PCA algorithm (PcaGrid) cutoff:
0.975
Choose an input file to upload
Browse...
No file selected
Analyze my data
Reset
Home
Input
Density Plot
hClustering
RobustPCA
Outliers
Rmarkdown
User Guide
EnsMOD: A Software Program for Omics Sample Outlier Detection
Detection of omics sample outliers is important for preventing erroneous biological conclusions, developing robust experimental protocols, and discovering rare biological states. Two recent publications describe robust algorithms for detecting transcriptomic sample outliers, but neither algorithm had been incorporated into a software tool for omics scientists. Ensemble Methods for Outlier Detection (EnsMOD) was developed by incorporating several different algorithms for robust analysis to identify potential outliers. EnsMOD calculates how closely the quantitation variation follows a normal distribution, plots the density curves of each sample to visualize anomalies, performs hierarchical cluster analyses to calculate how closely the samples cluster with each other, and performs robust principal component analyses to statistically test if any sample is an outlier. The probabilistic threshold parameters can be easily adjusted to tighten or loosen the outlier detection stringency. EnsMOD was used to analyze a simulated proteomics dataset, a multiomic (proteome and transcriptome) dataset, a single-cell proteomics dataset, and a phosphoproteomics dataset. EnsMOD successfully identified all of the simulated outliers, and outlier removal improved data quality for downstream statistical analyses.
Please refer to the User Guide on how to adjust the cutoff parameters, input file format, and understanding the results.
USA.gov Government Made Easy
Version 2.0, last update: 01/01/2023, Privacy Policy, Disclaimer, Vulnerability Disclosure Policy


EnsMOD - Ensemble Methods fo
127.0.0.1:6350
NIH
National Institute of Allergy and Infectious Diseases
Leading research to understand, treat, and prevent infectious, immunologic, and allergic diseases
EnsMOD - Ensemble Methods for Outlier Detection
Cophenetic Correlation Coefficient (CCC):
0.8
Silhouette Coefficient (SC):
0.25
Robust Sparse PCA algorithm (robpca) cutoff:
0.975
Robust PCA algorithm (PcaGrid) cutoff:
0.975
Choose an input file to upload
Browse...
Gene_Expression_Table.xlsx
Upload complete
Analyze my data
Reset
Home
Input
Density Plot
hClustering
RobustPCA
Outliers
Rmarkdown
User Guide
Show 10 entries
Search:
s_01 s_02 s_03 s_04 s_05 s_06 s_07
1 1743958.00091499 1607339.55191554 1756692.52148069 2101903.73348625 1749199.9954201 1275060.89345945 1483708.96515335
2 1143416.24190256 1260378.02081454 1210863.22550074 1383320.83440058 995030.612230222 707131.230177632 726080.491612985
3 953112.356597948 1057921.0026897 931752.339563426 893580.659363245 819710.532899739 1233818.70343106 1301985.55912744
4 893773.306229724 874166.173574901 889253.334481035 786868.751661702 951623.493822793 1343745.50439144 1445515.92289067
5 1315770.55620269 1179681.90160526 1544398.56762447 1299985.09627496 1181588.93977884 1167666.64018666 1244428.27477323
6 540462.123117657 588431.969060214 522673.637382686 538114.831535017 518881.655491043 1512276.11777871 1318667.58353497
7 1376938.06247903 1348523.66968152 1384685.91060681 1429206.40198935 1075386.29845986 1126789.8480047 1000493.27594756
8 464043.883024655 459124.328956373 482013.738516398 399809.683981754 464089.341795302 857804.338082964 741658.290695519
9 1215282.06151923 1291603.6823689 1191973.41873533 1342308.06180242 1227626.75600095 1283262.68234676 1324084.10423944
10 990069.116585969 1001884.31728915 953470.031268907 1083175.2853509 1114516.25967499 1312127.92164566 1288637.5360906
Showing 1 to 10 of 100 entries
Previous 1 2 3 4 5 ... 10 Next
USA.gov
Government Made Easy
Version 2.0, last update: 01/01/2023, Privacy Policy, Disclaimer, Vulnerability Disclosure Policy

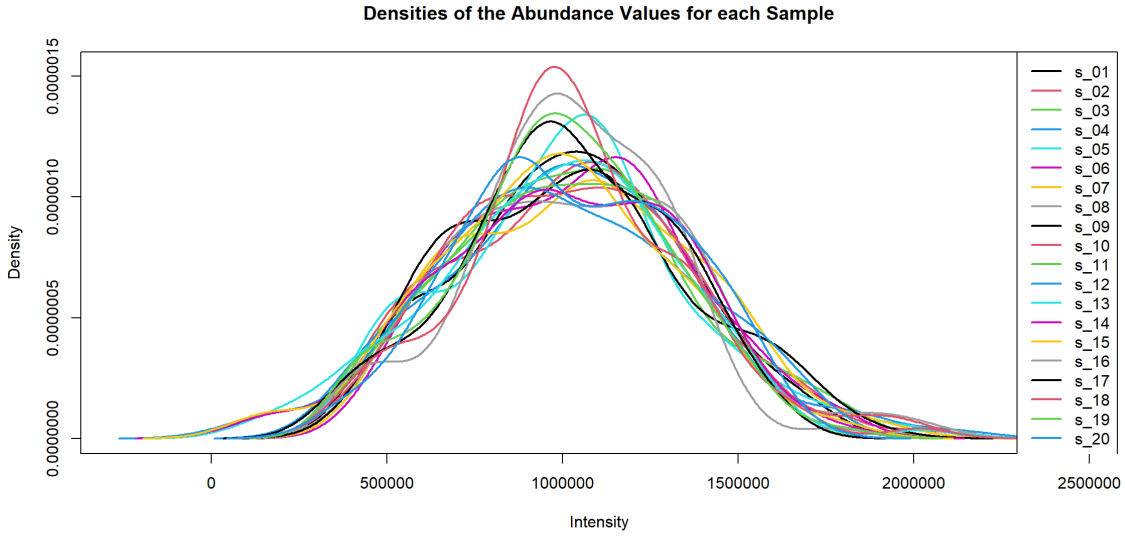
Interpretation of the EnsMOD Output

EnsMOD will produce an HTML output file that can be reviewed using a web browser. The output HTML file will be in the directory with the Rmarkdown in /app/www/EnsMODoutpupts/ (application version). All of the results are in the output HTML file, but the application version will also display individual results using the GUI, and it will also save these individual results as output files in the EnsMODoutpupts directory.

For this user guide, we used EnsMOD to analyze a simulated proteomics dataset (https://github.com/niaid/EnsMOD/tree/main/app/EnsMOD_Examples/Simulated_Proteomics_Data). There were twenty samples total, nineteen were partitioned into four groups (each contained four or five samples), and there was one outlier (sample s_20). For each protein and sample group, the abundance values were drawn from a normal distribution (coefficient of variation = 15%).

Density Curves.

A density curve for each sample was plotted.



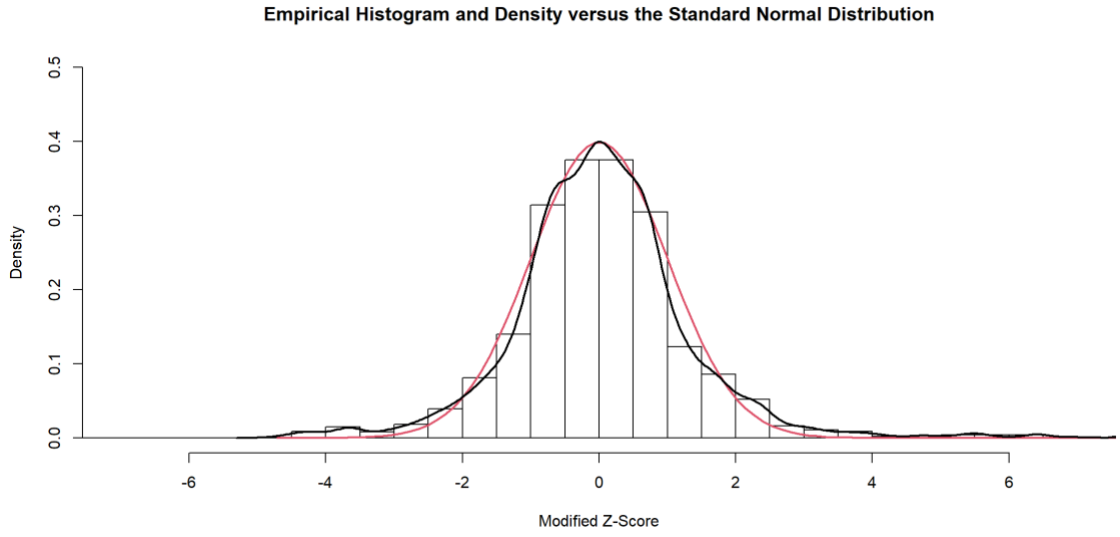
A sample with a density curve different from the others should be carefully investigated and might be an outlier.

Data Normality.

The rPCA outlier detection algorithms assume that the overall quantitation variation of the input dataset follows a normal distribution. No genuine experimental data variation exactly follows a normal distribution, and unfortunately it is unknown how non-normal the variation can be before

the rPCA outlier detection algorithms become erroneous. Therefore, EnsMOD includes multiple tools to inspect the normality of the variation.

An empirical histogram and density curve was plotted against a standard normal distribution (mean = 0, standard deviation = 1), and the corresponding coefficient of determination (R^2) was calculated (using the empirical density curve and the standard normal distribution).



Here, R-squared = 0.9863945. If the empirical histogram or density curve is skewed relative to the standard normal distribution, the data variance might be non-normal. Note that the modified Z-score is used. These values are robust to outliers but can cause artifacts. To use regular Z-scores, search for "To use regular Z-scores, enable the line below." in the EnsMOD Rmarkdown using RStudio and follow the instructions.

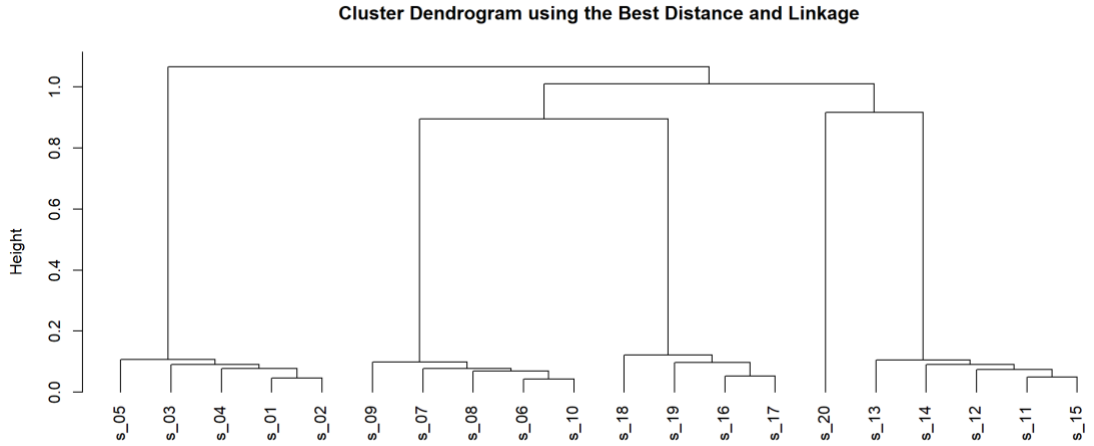
The empirical and fitted cumulative distributions are plotted, as are the Quantile-Quantile and Probability-Probability plots. The skewness versus kurtosis is plotted, and the Shapiro-Wilk and Kolmogorov-Smirnov tests for normality are performed. If the variation deviates too far from a normal distribution, then an upstream data transformation to achieve normality should be considered (Huber et al 2002; Kelmansky et al 2013; Raymaekers and Rousseeuw 2021; Rocke and Durbin 2003). Sometimes log-transformation of the data results in normally distributed variation.

Hierarchical Cluster Analysis

Fifteen hierarchical cluster analyses (HCAs) were performed using three distance functions (Euclidean, Manhattan, and Pearson) and five linkage functions (average, complete, single, centroid, and Ward.D2). For each HCA, the cophenetic correlation coefficient (CCC) was calculated and tabulated.

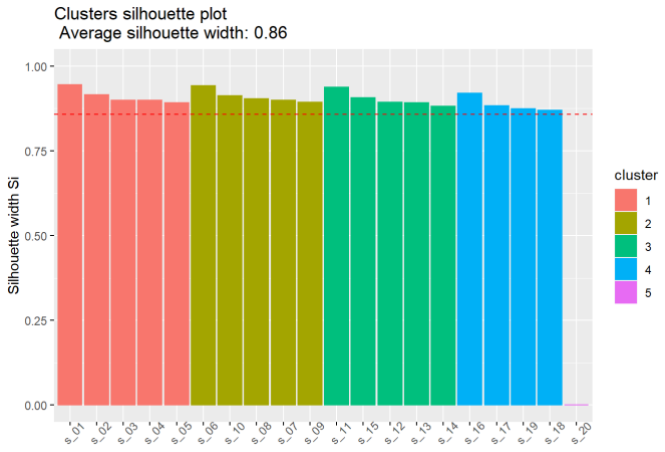
```
##     distance  linkage distance_matrix       CCC
## 11   pearson  average             p_a 0.9921174
## 14   pearson   single             p_s 0.9898400
## 1  euclidean  average             e_a 0.9871083
## 13   pearson complete            p_co 0.9859448
## 4  euclidean   single             e_s 0.9840743
## 6  manhattan  average             m_a 0.9835766
## 3  euclidean complete            e_co 0.9822055
## 9  manhattan   single             m_s 0.9803081
## 8  manhattan complete            m_co 0.9753873
## 12   pearson  ward.D2             p_w 0.9743125
## 2  euclidean  ward.D2             e_w 0.9711011
## 7  manhattan  ward.D2             m_w 0.9699020
## 15   pearson centroid            p_ce 0.9439930
## 10 manhattan centroid            m_ce 0.9091207
## 5  euclidean centroid            e_ce 0.9087703
```

The table was sorted by CCC (descending). The CCC is a measure of how well a dendrogram preserves the pairwise distances of the original dataset. The CCC can range from zero (the clustering is uninformative) to one (the clusters perfectly represent the original distances). A CCC ≥ 0.8 was required (this EnsMOD parameter can be adjusted) (Selicato et al 2021). If the CCC < 0.8, then the clustering is probably too poor to robustly identify outlier(s). The distance-linkage pair that resulted in the highest CCC was used for the downstream analyses.



While not statistically robust, it is clear from the dendrogram that sample s_20 (the simulated outlier) was the furthest from its nearest neighbor, and thus the most likely outlier.

The gap statistics algorithm was also used to calculate the optimal number of clusters (Selicato et al 2021). If this failed, the maximum average SC method was used (Charrad et al 2014). The Silhouette coefficient (SC) was calculated for each sample.



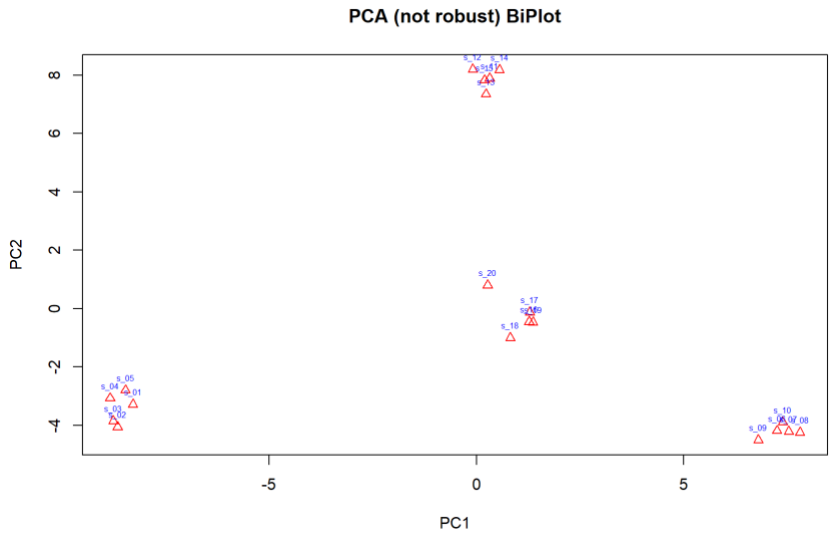

```
##      cluster neighbor sil_width
## s_01       1        3 0.9440836
## s_02       1        4 0.9149981
## s_03       1        4 0.8994732
## s_04       1        3 0.8984505
## s_05       1        5 0.8917571
## s_06       2        4 0.9421835
## s_10       2        4 0.9116910
## s_08       2        4 0.9032621
## s_07       2        4 0.8989609
## s_09       2        4 0.8929748
## s_11       3        5 0.9365698
## s_15       3        5 0.9068518
## s_12       3        5 0.8936783
## s_13       3        2 0.8911057
## s_14       3        5 0.8818247
## s_16       4        2 0.9200394
## s_17       4        2 0.8821357
## s_19       4        2 0.8741084
## s_18       4        2 0.8691730
## s_20       5        3 0.0000000
```

The SC can range from −1 (the sample would fit much better in a different cluster) through zero (the sample does not fit in any cluster) to one (the sample fits perfectly in its cluster). A sample with SC < 0.25 was considered a potential outlier (this EnsMOD parameter can be adjusted) ("No substantial structures have been found.") (Selicato et al 2021). In the above chart and table, only sample s_20 (the simulated outlier) satisfied SC < 0.25. Thus, only sample s_20 was classified as a potential outlier by the HCA analysis, and the other samples were classified as non-outliers. A sample with $0.25 \leq SC < 0.5$ would be borderline ("The structure is weak and may be artificial.") (Selicato et al 2021).

The SC for each sample was calculated using the *eclust()* function of the *factoextra* R package. Note that *eclust()* is limited to a maximum of twenty clusters. *eclust()* might work poorly for datasets with more than ~20 experimental conditions. Though not ideal, large omics datasets with more than ~20 experimental conditions could be analyzed by partitioning the samples into subsets (each having less than 20 experimental conditions) and using EnsMOD to analyze each subset separately (this would just be for the SC values; EnsMOD would be used normally for all of the other values).

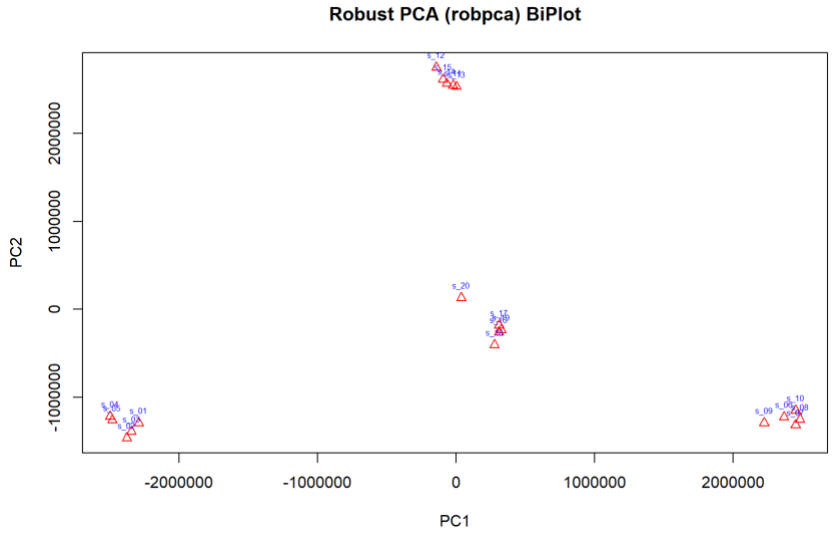
Robust Principal Component Analysis

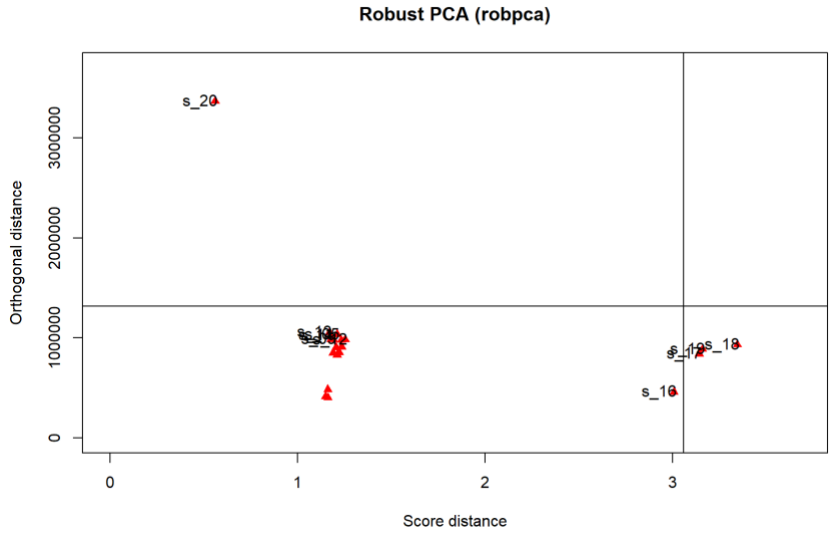
A classical principal component analysis (PCA) was performed to visualize the data on a biplot of principal component 1 versus principal component 2.

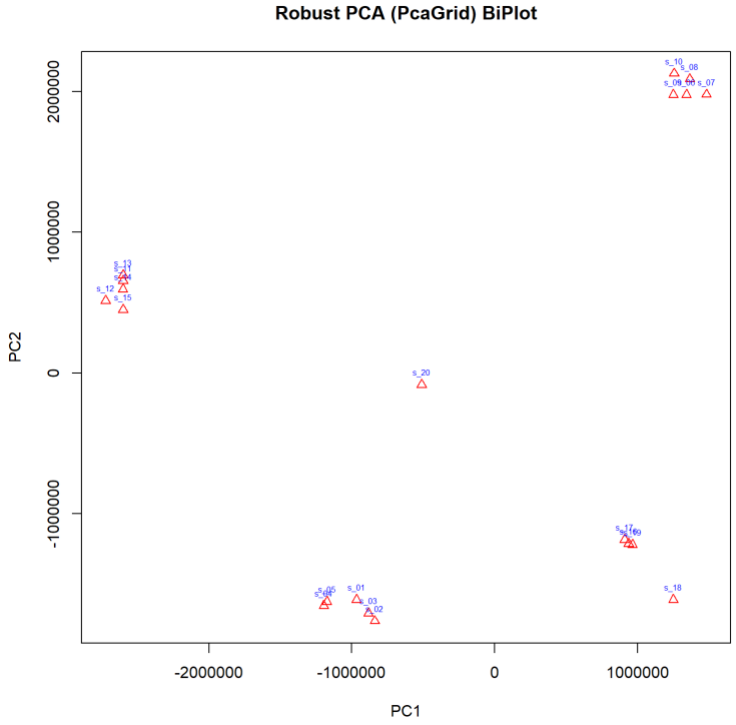


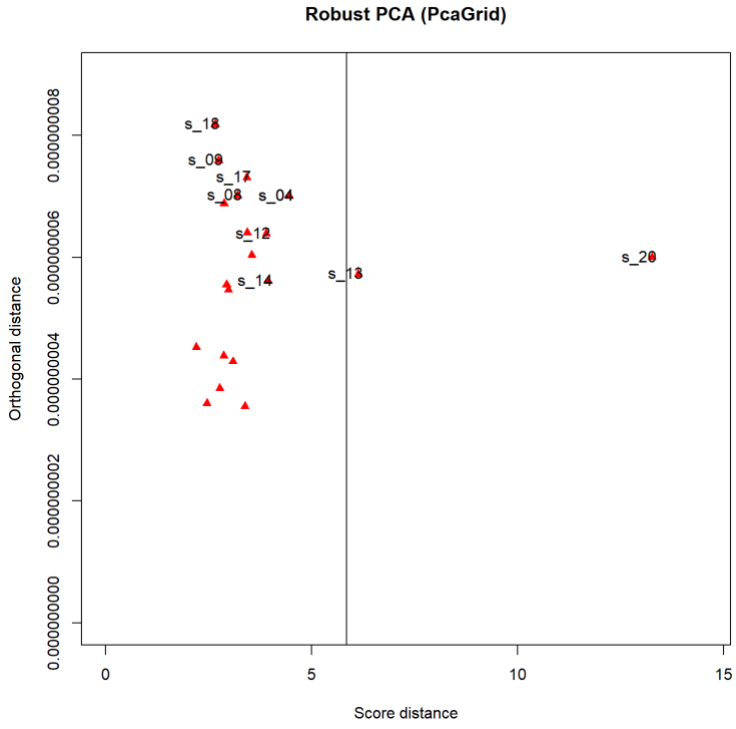
While not statistically robust, sample s_20 (the simulated outlier) was the furthest from its nearest neighbor, and thus the most likely outlier.

The ROBPCA (specifically, robpca of the rospca R package) and PcaGrid rPCA algorithms were performed, biplots of principal component 1 versus principal component 2 were made, and distance-distance plots were used to robustly detect outliers.






Robust PCA (PcaGrid) BiPlot
PC2
PC1
2000000
1000000
0
-1000000
-2000000
-1000000
0
1000000


Robust PCA (PcaGrid)
Orthogonal distance
Score distance
0.000000008
0.000000006
0.000000004
0.000000002
0.000000000
0
5
10
15
s_18
s_09
s_17
s_08
s_04
s_12
s_14
s_13
s_20

The vertical and horizontal lines in the distance-distance plots depict the stringency parameters (discussed below). Any sample that is within both thresholds (i.e., located within the lower-left region) was classified by the rPCA as a non-outlier, and the other samples were classified as outliers. The classification of each sample was tabulated:

```
# Display the robpca results (outliers are 'FALSE')
resR0_flag <- as.data.frame(resR0$flag.all)
resR0_flag
```

```
##      resR0$flag.all
## s_01           TRUE
## s_02           TRUE
## s_03           TRUE
## s_04           TRUE
## s_05           TRUE
## s_06           TRUE
## s_07           TRUE
## s_08           TRUE
## s_09           TRUE
## s_10           TRUE
## s_11           TRUE
## s_12           TRUE
## s_13           TRUE
## s_14           TRUE
## s_15           TRUE
## s_16           TRUE
## s_17          FALSE
## s_18          FALSE
## s_19          FALSE
## s_20          FALSE
```

```
# Display the results (outliers = FALSE)
pc_flag <- as.data.frame(pc$flag)
pc_flag
```

```
##      pc$flag
## s_01    TRUE
## s_02    TRUE
## s_03    TRUE
## s_04    TRUE
## s_05    TRUE
## s_06    TRUE
## s_07    TRUE
## s_08    TRUE
## s_09    TRUE
## s_10    TRUE
## s_11    TRUE
## s_12    TRUE
## s_13   FALSE
## s_14    TRUE
## s_15    TRUE
## s_16    TRUE
## s_17    TRUE
## s_18    TRUE
## s_19    TRUE
## s_20   FALSE
```

The PcaGrid outlier detection stringency parameter was set to 97.5% (for both the score distance test and the orthogonal distance test) (this EnsMOD parameter can be adjusted) (Chen et al 2020). For each test, and for data with normally distributed variation, this value is the estimated fraction of the samples that are not falsely classified as outliers. Likewise, robpca was used with the outlier detection stringency parameter set to 97.5% (for both the score distance test and the orthogonal distance test) (this EnsMOD parameter can be adjusted) (Selicato et al 2021).

We recommend considering the use of relatively strict rPCA thresholds with omics datasets (rPCA threshold of 0.99 or 0.999 seemed to work well with some of the example datasets). We also recommend using all four criteria (CCC, SC, robpca, and PcaGrid) for outlier detection. We recommend against using robpca alone.

Summary

At the end of the EnsMOD HTML output, the results are summarized.

5.1 Outlier(s) Identified by Robust PCA analyses

```
# A robpca cutoff of 0.975 is recommended (Selicato et al 2021).
# At the start of the script, the robpca cutoff was set to:
robpca_prob
```

```
## [1] 0.975
```

```
# The samples that are outside this cutoff are potential outliers, and the other samples are not.
# The samples that are outside this cutoff:
rosOutliers
```

```
## [1] "s_17" "s_18" "s_19" "s_20"
```

```
# A PcaGrid cutoff of 0.975 is recommended (Chen et al 2020).
# At the start of the script, the PcaGrid cutoff was set to:
PcaGrid_prob
```

```
## [1] 0.975
```

```
# The samples that are outside this cutoff are potential outliers, and the other samples are not.
# The samples that are outside this cutoff:
pcOutliers
```

```
## [1] "s_13" "s_20"
```

```
# The samples that satisfied both robust PCA (robpca and PcaGrid) criteria for an outlier:
intersect(pcOutliers, rosOutliers)
```

```
## [1] "s_20"
```

5.2 Outlier(s) Identified by Hiearchical Clustering

```
# A CCC cutoff of 0.8 is recommended (Selicato et al 2021).
# At the start of the script, the CCC cutoff was set to:
CCC_min
```

```
## [1] 0.8
```

```
# The CCC was calculated:
CCC_df_ranked_top$CCC
```

```
## [1] 0.9921174
```

```
# Did the input data pass the CCC test? (TRUE = yes, FALSE = no)
# If TRUE, then the HCA clustering was informative, and subsequent outlier detection can be informative.
# If FALSE, then the HCA clustering wasn't informative, and subsequent outlier detection won't be informative.
CCC_df_ranked_top$CCC >= CCC_min
```

```
## [1] TRUE
```

```
# An SC cutoff of 0.25 is recommended (Selicato et al 2021).
# At the start of the script, the SC cutoff was set to:
SC_max
```

```
## [1] 0.25
```

```
# The samples that have an SC value lower than the SC cutoff
# are potential outliers, and the other samples are not.
# The samples that have an SC value lower than the SC cutoff:
hcOutliers
```

```
## [1] "s_20"
```

5.3 Outlier(s) Identified by Both HCA and Robust PCA

```
# The samples that satisfied all four criteria (CCC, SC, robpca, PcaGrid) for an outlier:
if (CCC_df_ranked_top$CCC >= CCC_min) {
  intersect(intersect(hcOutliers, rosOutliers), pcOutliers)
}
```

```
## [1] "s_20"
```

References

Charrad M, Ghazzali N, Boiteau V et al. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. Journal of Statistical Software 2014; 61:1-36; doi: 10.18637/jss.v061.i06

Chen X, Zhang B, Wang T et al. Robust principal component analysis for accurate outlier sample detection in RNA-Seq data. BMC Bioinformatics 2020; 21:269; doi: 10.1186/s12859-020-03608-0

Huber W, von Heydebreck A, Sultmann H et al. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 2002; 18 Suppl 1:S96-104; doi: 10.1093/bioinformatics/18.suppl_1.s96

Kelmansky DM, Martinez EJ, Leiva V. A new variance stabilizing transformation for gene expression data analysis. Stat Appl Genet Mol Biol 2013; 12:653-66; doi: 10.1515/sagmb-2012-0030

Raymaekers J, Rousseeuw PJ. Transforming variables to central normality. Machine Learning 2021; doi: 10.1007/s10994-021-05960-5

Rocke DM, Durbin B. Approximate variance-stabilizing transformations for gene-expression microarray data. Bioinformatics 2003; 19:966-72; doi: 10.1093/bioinformatics/btg107

Selicato L, Esposito F, Gargano G et al. A New Ensemble Method for Detecting Anomalies in Gene Expression Matrices. Mathematics 2021; 9:882; doi: 10.3390/math9080882