

Ensemble Methods for Outlier Detection (EnsMOD): A Software Program for Omics Sample Outlier Detection

June 1, 2023

Open camera or QR reader and
scan code to access this article
and other resources online.



EnsMOD: A Software Program for Omics Sample Outlier Detection

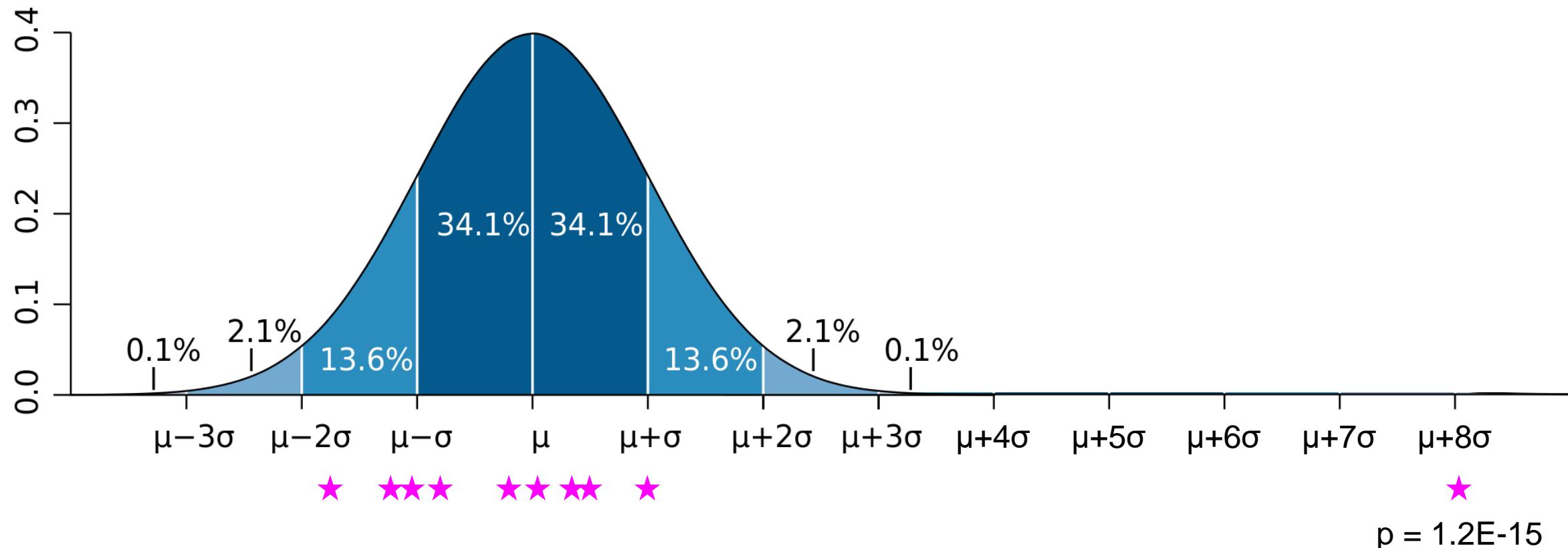
NATHAN P. MANES,* JIAN SONG,* and ALEKSANDRA NITA-LAZAR

ABSTRACT

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 scientists. Here we describe Ensemble Methods for Outlier Detection (EnsMOD) which incorporates both algorithms. 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 can be used to analyze any omics dataset with normally distributed variance. Here it 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 subsequent removal of a detected outlier improved data quality for downstream statistical analyses.

An Outlier is a Statistical Anomaly

“An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism.”



Outlier Mitigation Strategies

- Do nothing, but this might skew downstream statistics
- Increase the number of samples to wash out the effect of the outlier(s)
- Redo the entire experiment from the beginning
- Identify the outlier mechanism (if it is reproducible)
 - Biological outlier: Identify a rare biological mechanism
 - Technical outlier: Improve experimental protocol to prevent future outliers
- Outlier winsorizing (attenuation of extreme values)
- Outlier removal

Examples of Proteomics Sample Outliers

Example 1: LC-MS of a tissue proteomics sample, but the vial had an air bubble resulting in an unintended blank run

Result: Tissue proteins missing

Example 2: A surgeon prepares 20 clinical tissue biopsy samples, but one originated from a nearby region

Result: Tissue proteins have a different pattern

Two Recent Articles Describe Robust Transcriptomics Sample Outlier Detection

Chen et al. BMC Bioinformatics (2020) 21:269
<https://doi.org/10.1186/s12859-020-03608-0>

BMC Bioinformatics

RESEARCH ARTICLE

Open Access



Check for
updates

Robust principal component analysis for accurate outlier sample detection in RNA-Seq data

Xiaoying Chen¹, Bo Zhang², Ting Wang^{3,4}, Azad Bonni¹ and Guoyan Zhao^{1*}

* Correspondence: gzhao@wustl.edu

¹Department of Neuroscience, Washington University School of Medicine, St. Louis, MO, USA
Full list of author information is available at the end of the article

Abstract

Background: High throughput RNA sequencing is a powerful approach to study gene expression. Due to the complex multiple-steps protocols in data acquisition, extreme deviation of a sample from samples of the same treatment group may occur due to technical variation or true biological differences. The high-dimensionality of the data with few biological replicates make it challenging to accurately detect those samples, and this issue is not well studied in the literature currently. Robust statistics is a family of theories and techniques aim to detect the outliers by first fitting the majority of the data and then flagging data points that deviate from it. Robust statistics have been widely used in multivariate data analysis for outlier detection in chemometrics and engineering. Here we apply robust statistics on RNA-seq data analysis.

Results: We report the use of two robust principal component analysis (rPCA) methods, *PcaHubert* and *PcaGrid*, to detect outlier samples in multiple simulated and real biological RNA-seq data sets with positive control outlier samples. *PcaGrid* achieved 100% sensitivity and 100% specificity in all the tests using positive control outliers with varying degrees of divergence. We applied rPCA methods and classical principal component analysis (cPCA) on an RNA-Seq data set profiling gene expression of the external granule layer in the cerebellum of control and conditional *SnoN* knockout mice. Both rPCA methods detected the same two outlier samples but cPCA failed to detect any. We performed differentially expressed gene detection before and after outlier removal as well as with and without batch effect modeling. We validated gene expression changes using quantitative reverse transcription PCR and used the result as reference to compare the performance of eight different data

mathematics



Article

A New Ensemble Method for Detecting Anomalies in Gene Expression Matrices

Laura Selicato ^{1,2,*†} , Flavia Esposito ^{1,2,†} , Grazia Gargano ¹ , Maria Carmela Vegliante ³ , Giuseppina Opinto ³ , Gian Maria Zaccaria ³ , Sabino Ciavarella ³ , Attilio Guarini ³ , and Nicoletta Del Buono ^{1,2}

¹ Department of Mathematics, University of Bari Aldo Moro, 70125 Bari, Italy; flavia.esposito@uniba.it (F.E.); g.gargano20@studenti.uniba.it (G.G.); nicoletta.delbuono@uniba.it (N.D.B.)

² Member of GNCS, Istituto Nazionale di Alta Matematica, Ple Aldo Moro 5, 00185 Roma, Italy

³ Hematology and Cell Therapy Unit, IRCCS-Istituto Tumori 'Giovanni Paolo II', 70124 Bari, Italy; mc.vegliante@oncologico.bari.it (M.C.V.); giusyopinto@hotmail.it (G.O.); gianmari.zaccaria@gmail.com (G.M.Z.); sabino.ciavarella@yahoo.it (S.C.); attilio.guarini@oncologico.bari.it (A.G.)

* Correspondence: laura.selicato@uniba.it

† These authors contributed equally to this work.



check for
updates

Citation: Selicato, L.; Esposito, F.; Gargano, G.; Vegliante, M.C.; Opinto, G.; Zaccaria, G.M.; Ciavarella, S.; Guarini, A.; Del Buono, N. A New Ensemble Method for Detecting Anomalies in Gene Expression Matrices. *Mathematics* **2021**, *9*, 882. <https://doi.org/10.3390/math9080882>

Academic Editor: Junseok Kim

Received: 1 March 2021

Accepted: 14 April 2021

Abstract: One of the main problems in the analysis of real data is often related to the presence of anomalies. Namely, anomalous cases can both spoil the resulting analysis and contain valuable information at the same time. In both cases, the ability to detect these occurrences is very important. In the biomedical field, a correct identification of outliers could allow the development of new biological hypotheses that are not considered when looking at experimental biological data. In this work, we address the problem of detecting outliers in gene expression data, focusing on microarray analysis. We propose an ensemble approach for detecting anomalies in gene expression matrices based on the use of Hierarchical Clustering and Robust Principal Component Analysis, which allows us to derive a novel pseudo-mathematical classification of anomalies.

Keywords: anomaly; low rank decomposition; gene expression; clustering; outliers

1. Introduction

Real datasets often contain observations that behave differently from the majority of the data. If an occurrence differs from the dominant part of the data, or if it is sufficiently unlikely under the assumed data probability model, it is considered an anomaly or outlier

but no software tool had been developed for omics scientists

EnsMOD: A Software Program for Omics Sample Outlier Detection

The screenshot shows the GitHub repository page for EnsMOD. The repository is public and has 92 commits. The repository description is: "EnsMOD: A Software Program for Omics Sample Outlier Detection". It includes sections for About, Releases, Packages, Contributors, and Languages. The Languages section shows HTML at 99.8% and Other at 0.2%.

About
EnsMOD: A Software Program for Omics Sample Outlier Detection

Releases
No releases published

Packages
No packages published

Contributors 2

- jiansongatnih Jian
- manesnp

Languages

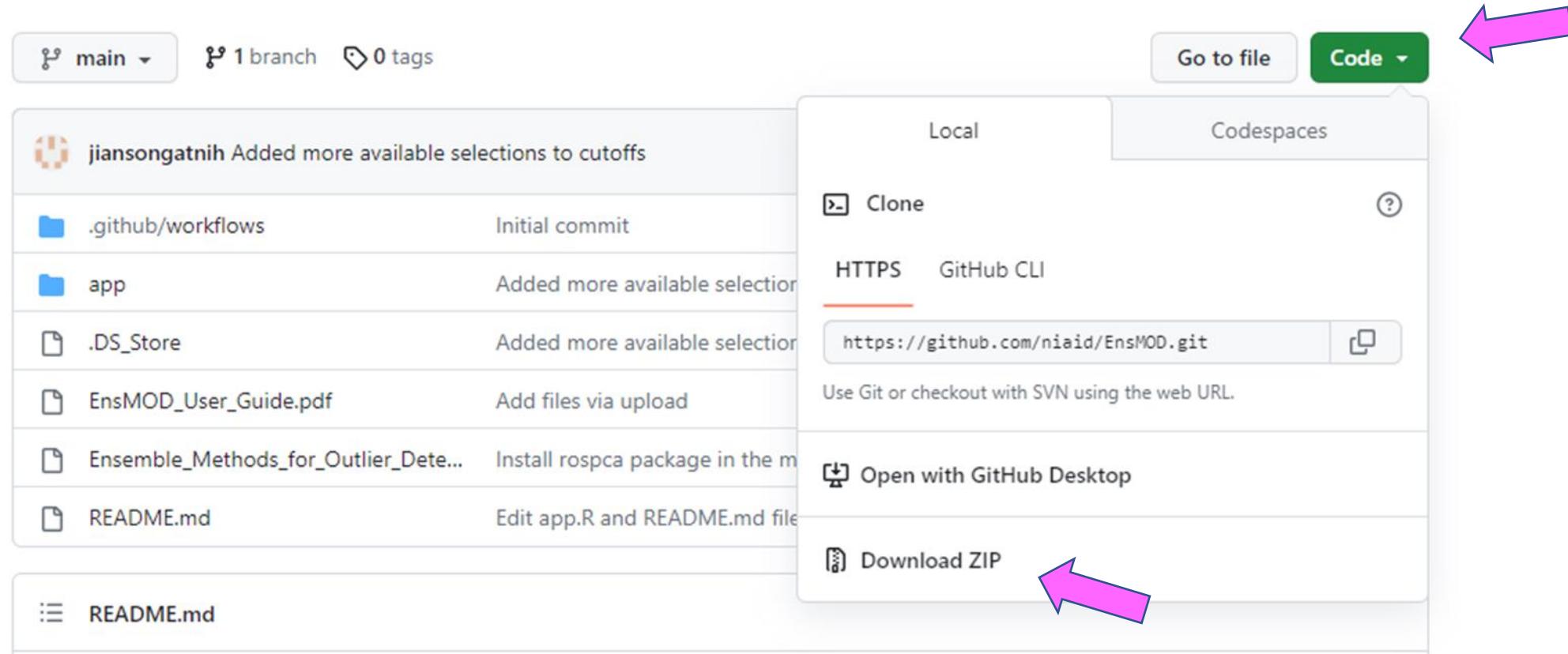
- HTML 99.8%
- Other 0.2%

- Free and open-source
- Uses both previously published algorithms for robust transcriptomics sample outlier detection
- Can be used to analyze any omics dataset (~Gaussian variance, ≥ 9 samples)
- No need for a strong background in bioinformatics or biostatistics
- Step-by-step user guide
- Published

<https://github.com/niaid/EnsMOD>

Using EnsMOD

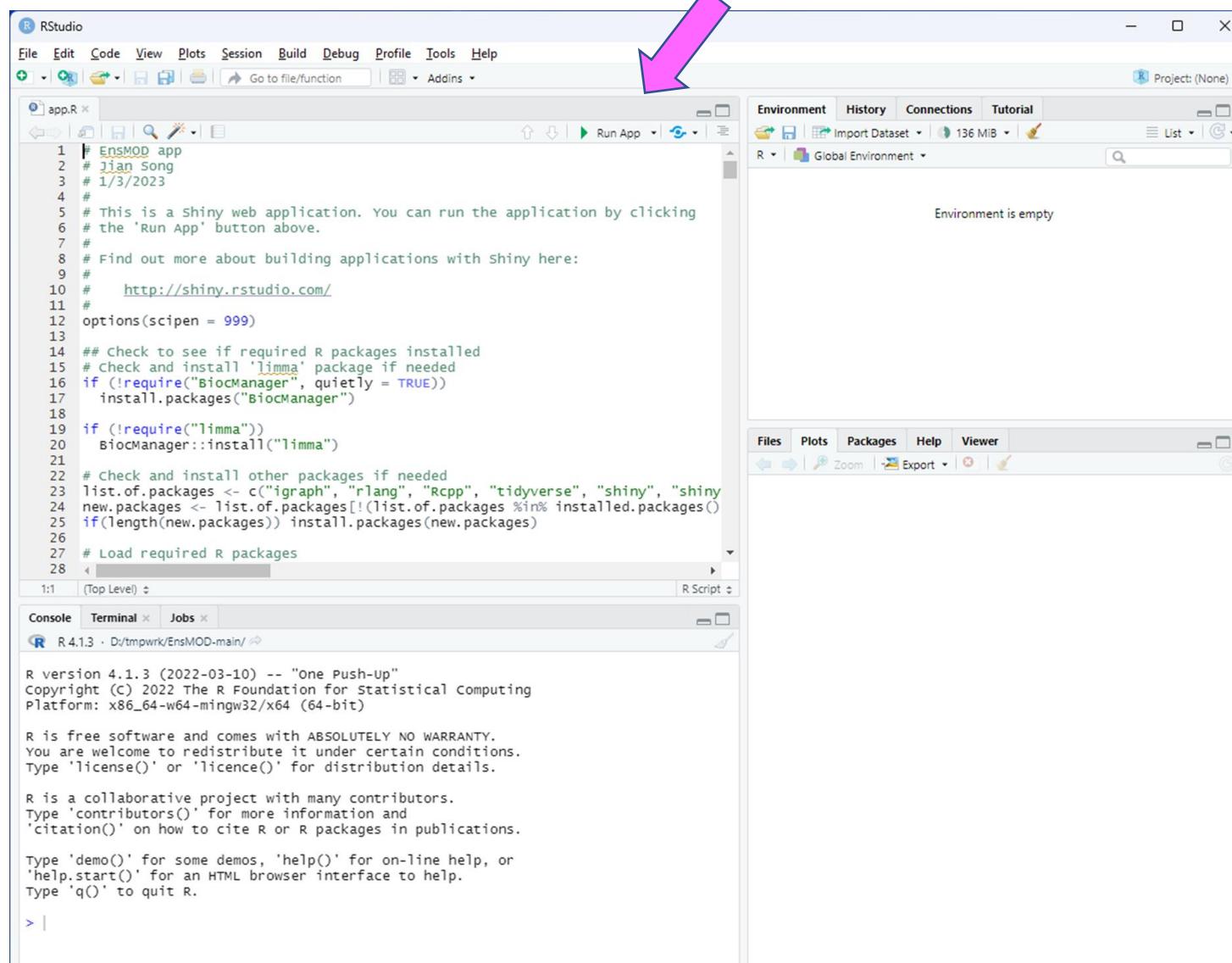
1. Install R from <https://www.r-project.org/>
2. Install RStudio from <https://www.rstudio.com/>
3. Download and unzip EnsMOD: <https://github.com/niaid/EnsMOD>



Using EnsMOD

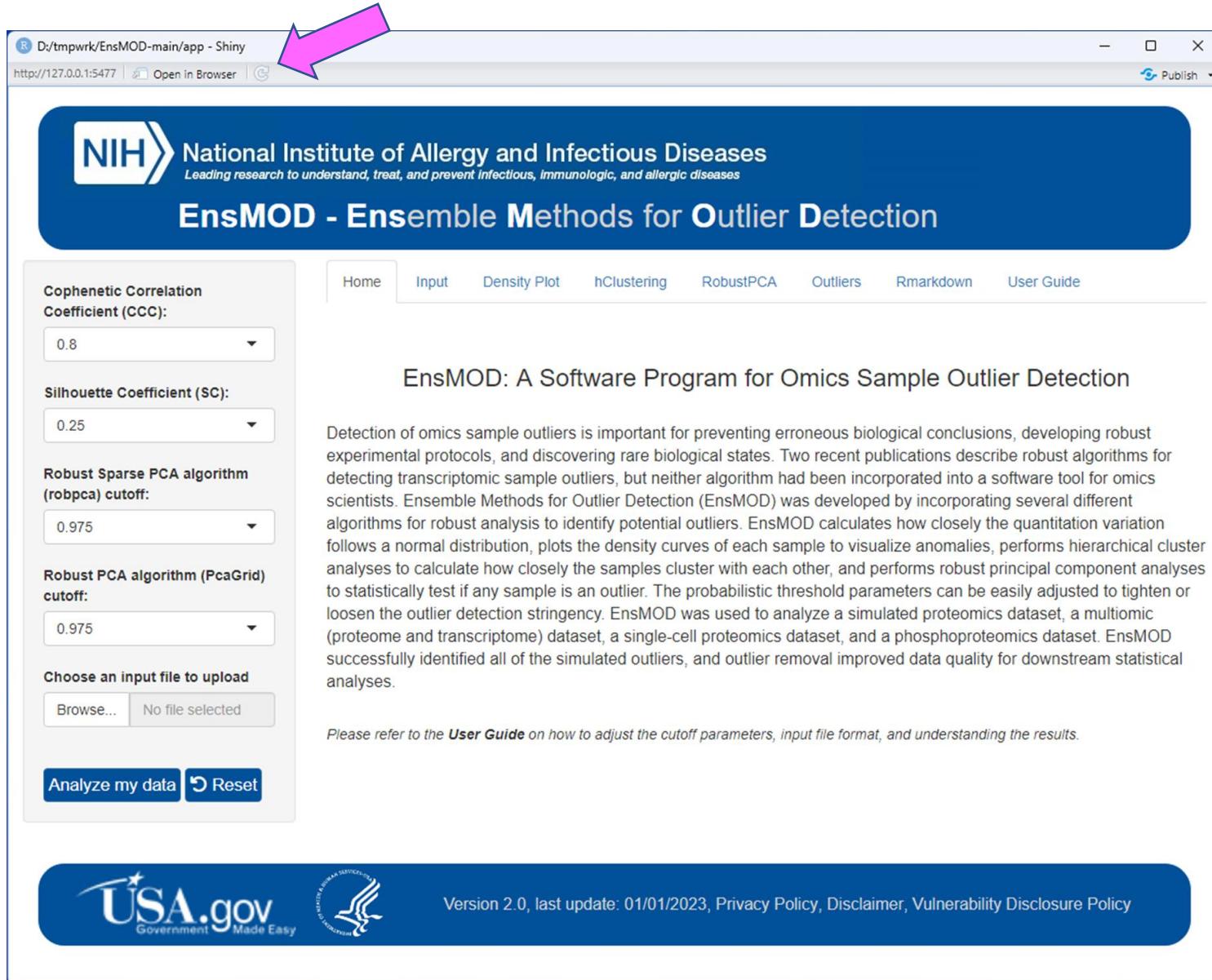
4. Run RStudio and open `/app/app.R`

5. Click “Run App”



Using EnsMOD

6. Click “Open in Browser” (optional)



The screenshot shows the EnsMOD Shiny application running in a web browser. The title bar indicates the path is D:/tmpwrk/EnsMOD-main/app - Shiny and the URL is http://127.0.0.1:5477. A pink arrow points to the "Open in Browser" button in the top left corner of the browser window.

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):

Silhouette Coefficient (SC):

Robust Sparse PCA algorithm (robPCA) cutoff:

Robust PCA algorithm (PcaGrid) cutoff:

Choose an input file to upload

Analyze my data

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.

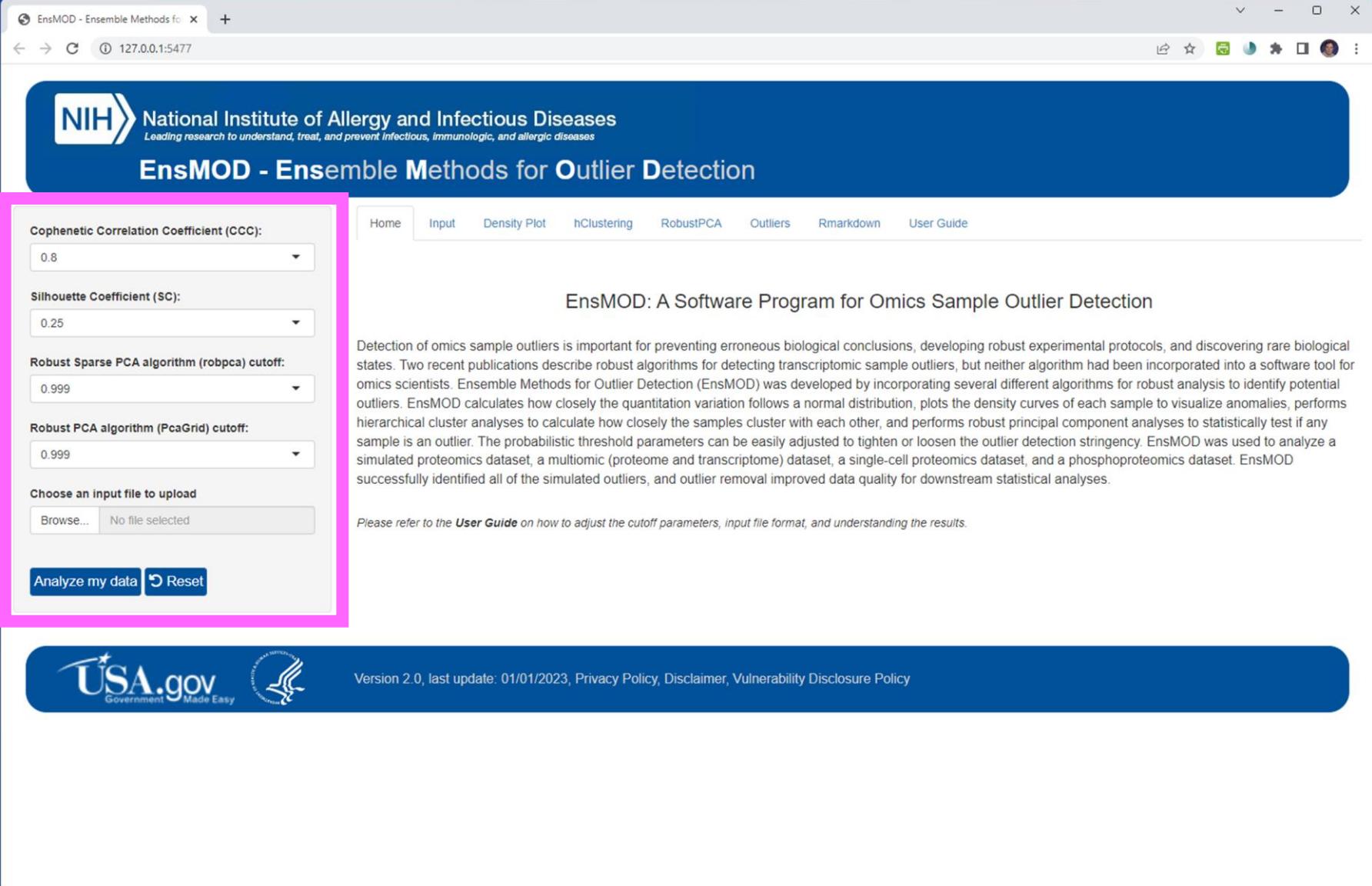
Please refer to the [User Guide](#) on how to adjust the cutoff parameters, input file format, and understanding the results.

USA.gov Government Made Easy

Version 2.0, last update: 01/01/2023, [Privacy Policy](#), [Disclaimer](#), [Vulnerability Disclosure Policy](#)

Using EnsMOD

7. Set parameters and select expression file



The screenshot shows the EnsMOD web application running in a browser window. The title bar reads "EnsMOD - Ensemble Methods fo 127.0.0.1:5477". The header features the NIH logo and the text "National Institute of Allergy and Infectious Diseases" with the subtitle "Leading research to understand, treat, and prevent infectious, immunologic, and allergic diseases". Below the header is the main title "EnsMOD - Ensemble Methods for Outlier Detection".

The left side of the interface contains a form with several dropdown menus and a file upload field:

- Cophenetic Correlation Coefficient (CCC): 0.8
- Silhouette Coefficient (SC): 0.25
- Robust Sparse PCA algorithm (robPCA) cutoff: 0.999
- Robust PCA algorithm (PcaGrid) cutoff: 0.999
- Choose an input file to upload: No file selected

Below the form are two buttons: "Analyze my data" and "Reset".

The right side of the page contains descriptive text about EnsMOD's purpose and capabilities:

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.

At the bottom of the page are the USA.gov "Government Made Easy" logo and a link to the "Version 2.0, last update: 01/01/2023, Privacy Policy, Disclaimer, Vulnerability Disclosure Policy".

EnsMOD Input Expression File

Input file format: CSV, XLSX, or TXT (tab delimited)

Columns = samples

Rows = analytes

Values = analyte abundance values (~Gaussian variance, ≥ 9 samples)

First row = sample unique identifiers

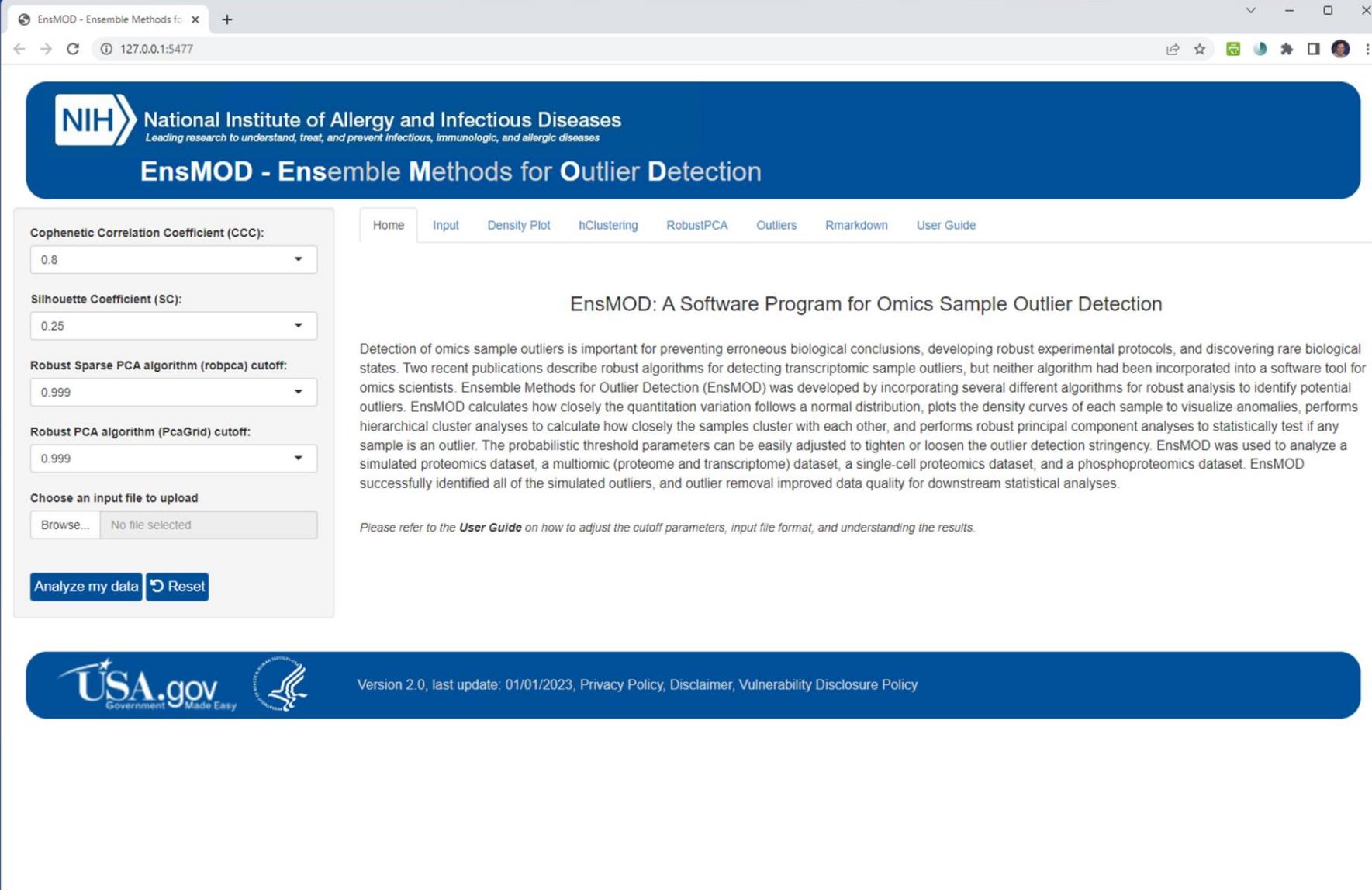
All columns are of abundance data (no column of analyte unique identifiers)

Note: Rows that contain ≥ 1 missing values (e.g., “NaN”) are ignored

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	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	
2	1743958	1607340	1756693	2101904	1749200	1275061	1483709	1438430	1207972	1306720	1166540	934368.6	1076671	1168147	1035184	1197665	1362627	1155776	1268309	868575.7	
3	1143416	1260378	1210863	1383321	995030.6	707131.2	726080.5	620831.6	827087.8	702240.7	1453863	1428003	1398112	1287705	1370912	1279348	1217641	1354420	1177517	1314460	
4	953112.4	1057921	931752.3	893580.7	819710.5	1233819	1301986	1337808	1187208	1125274	815100.2	848613.9	883181	951964.9	895318.1	1130952	1078530	1034869	1185874	748566.6	
5	893773.3	874166.2	889253.3	786868.8	951623.5	1343746	1445516	1319454	1614338	1455375	941261.6	950279.1	1001413	1094408	897009.9	1037834	987344.4	939823.8	1027515	1366319	
6	1315771	1179682	1544399	1299985	1181589	1167667	1244428	1200741	993969.4	1239823	1234697	1458223	1420999	1195886	1018476	1219898	1338259	983564.2	1098745	1110333	
7	540462.1	588432	522673.6	538114.8	518881.7	1512276	1318668	1604725	1448653	1562603	1516618	1390069	1500380	1602273	1378565	930487.3	927418.9	960633	937892.9	833902.6	

Using EnsMOD

7. Set parameters and select expression file



The image shows the EnsMOD software interface running in a web browser. The title bar indicates the URL is 127.0.0.1:5477. The header features the NIH logo and the text "National Institute of Allergy and Infectious Diseases" with the subtitle "Leading research to understand, treat, and prevent infectious, immunologic, and allergic diseases". Below the header, the main title "EnsMOD - Ensemble Methods for Outlier Detection" is displayed.

The left sidebar contains four dropdown menus for setting parameters:

- Cophenetic Correlation Coefficient (CCC): 0.8
- Silhouette Coefficient (SC): 0.25
- Robust Sparse PCA algorithm (robPCA) cutoff: 0.999
- Robust PCA algorithm (PcaGrid) cutoff: 0.999

Below these parameters is a section titled "Choose an input file to upload" with a "Browse..." button and a "No file selected" label. At the bottom of this sidebar are two buttons: "Analyze my data" and "Reset".

The right side of the interface displays a summary of the software's purpose and capabilities:

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.

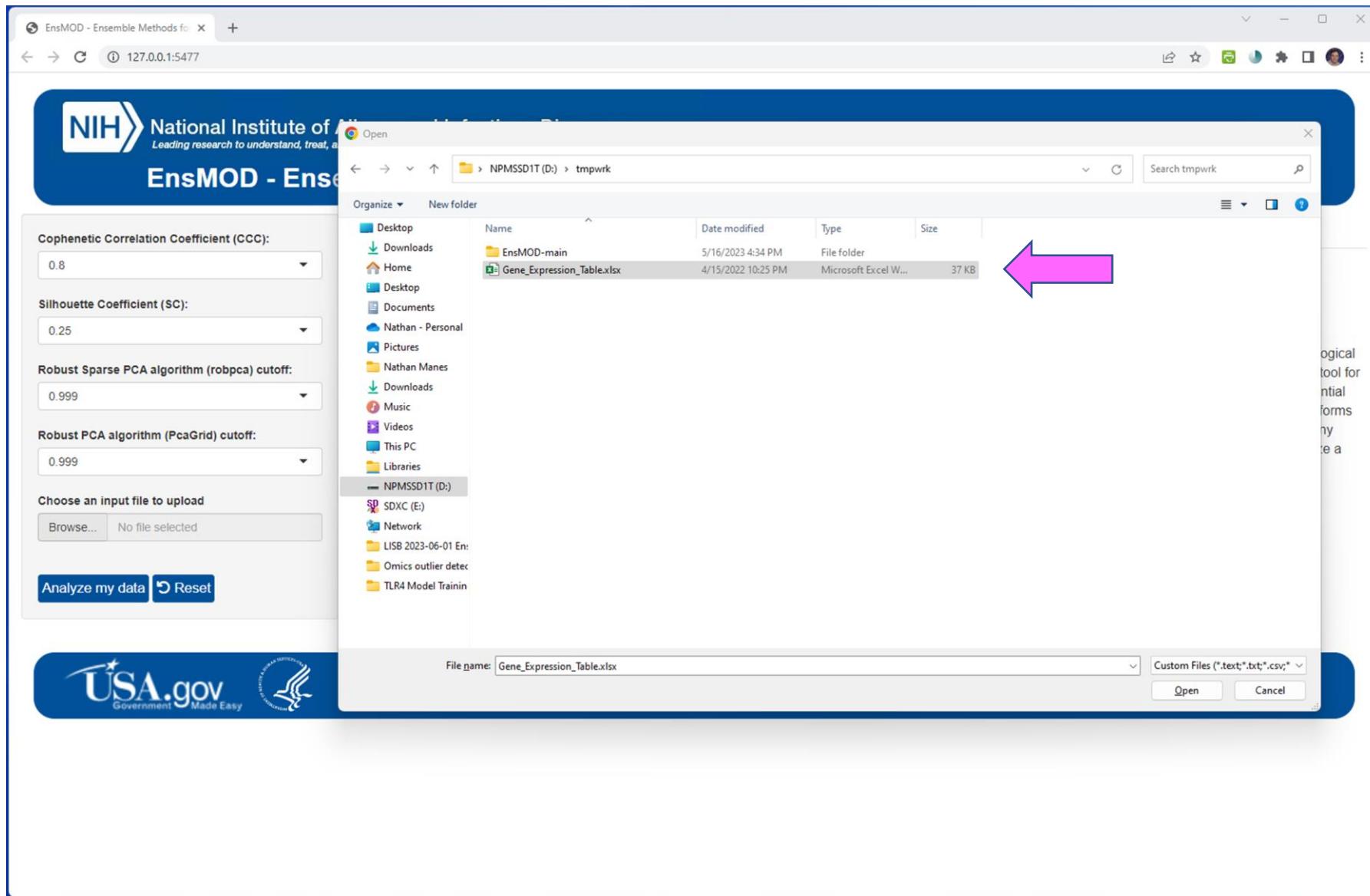
Please refer to the [User Guide](#) on how to adjust the cutoff parameters, input file format, and understanding the results.

At the bottom of the page, there is a USA.gov "Government Made Easy" logo and a link to the "Version 2.0, last update: 01/01/2023, Privacy Policy, Disclaimer, Vulnerability Disclosure Policy".

A large pink arrow points to the "Choose an input file to upload" field on the left side of the interface.

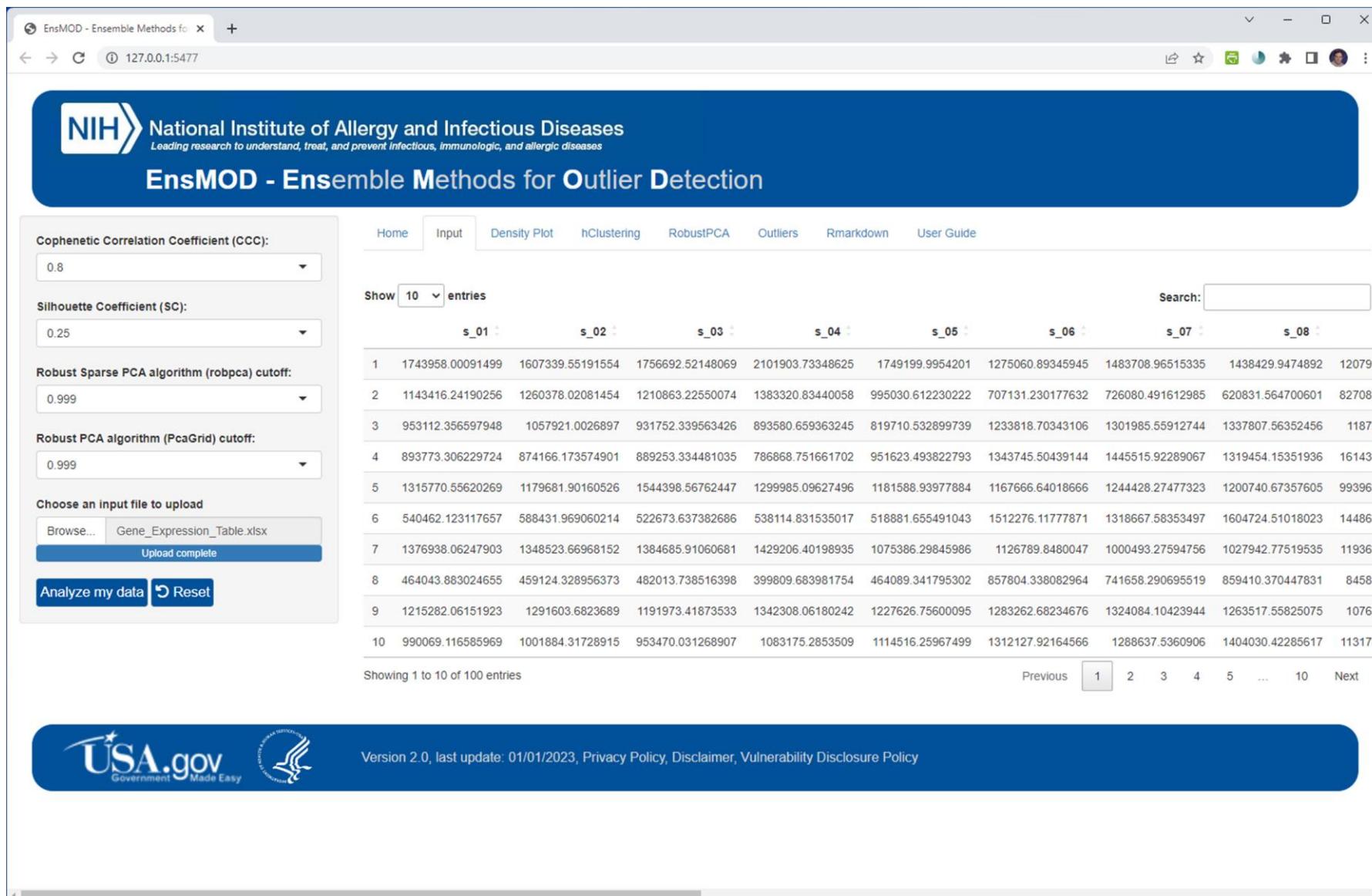
Using EnsMOD

7. Set parameters and select expression file



Using EnsMOD

8. Click “Analyze my data”



The screenshot shows the EnsMOD - Ensemble Methods for Outlier Detection web application. The interface includes:

- NIH National Institute of Allergy and Infectious Diseases logo** and **Leading research to understand, treat, and prevent infectious, immunologic, and allergic diseases** tagline.
- EnsMOD - Ensemble Methods for Outlier Detection** title.
- Input parameters:**
 - Cophenetic Correlation Coefficient (CCC): 0.8
 - Silhouette Coefficient (SC): 0.25
 - Robust Sparse PCA algorithm (robPCA) cutoff: 0.999
 - Robust PCA algorithm (PcaGrid) cutoff: 0.999
- File upload:** Choose an input file to upload, showing "Gene_Expression_Table.xlsx" and "Upload complete".
- Buttons:** "Analyze my data" (highlighted by a pink arrow) and "Reset".
- Data table:** Shows 10 entries of data with columns labeled s_01 through s_08. The data is as follows:

	s_01	s_02	s_03	s_04	s_05	s_06	s_07	s_08	
1	1743958.00091499	1607339.55191554	1756692.52148069	2101903.73348625	1749199.9954201	1275060.89345945	1483708.96515335	1438429.9474892	1207971
2	1143416.24190256	1260378.02081454	1210863.22550074	1383320.83440058	995030.612230222	707131.230177632	726080.491612985	620831.564700601	8270871
3	953112.356597948	1057921.0026897	931752.339563426	893580.659363245	819710.532899739	1233818.70343106	1301985.55912744	1337807.56352456	118720
4	893773.306229724	874166.173574901	889253.334481035	786868.751661702	951623.493822793	1343745.50439144	1445515.92289067	1319454.15351936	1614338
5	1315770.55620269	1179681.90160526	1544398.56762447	1299985.09627496	1181588.93977884	1167666.64018666	1244428.27477323	1200740.67357605	9939691
6	540462.123117657	588431.969060214	522673.637382686	538114.831535017	518881.655491043	1512276.11777871	1318667.58353497	1604724.51018023	1448652
7	1376938.06247903	1348523.66968152	1384685.91060681	1429206.40198935	1075386.29845986	1126789.8480047	1000493.27594756	1027942.77519535	1193690
8	464043.883024655	459124.328956373	482013.738516398	399809.683981754	464089.341795302	857804.338082964	741658.290695519	859410.370447831	845845
9	1215282.06151923	1291603.6823689	1191973.41873533	1342308.06180242	1227626.75600095	1283262.68234676	1324084.10423944	1263517.55825075	107654
10	990069.116585969	1001884.31728915	953470.031268907	1083175.2853509	1114516.25967499	1312127.92164566	1288637.5360906	1404030.42285617	1131784

Showing 1 to 10 of 100 entries

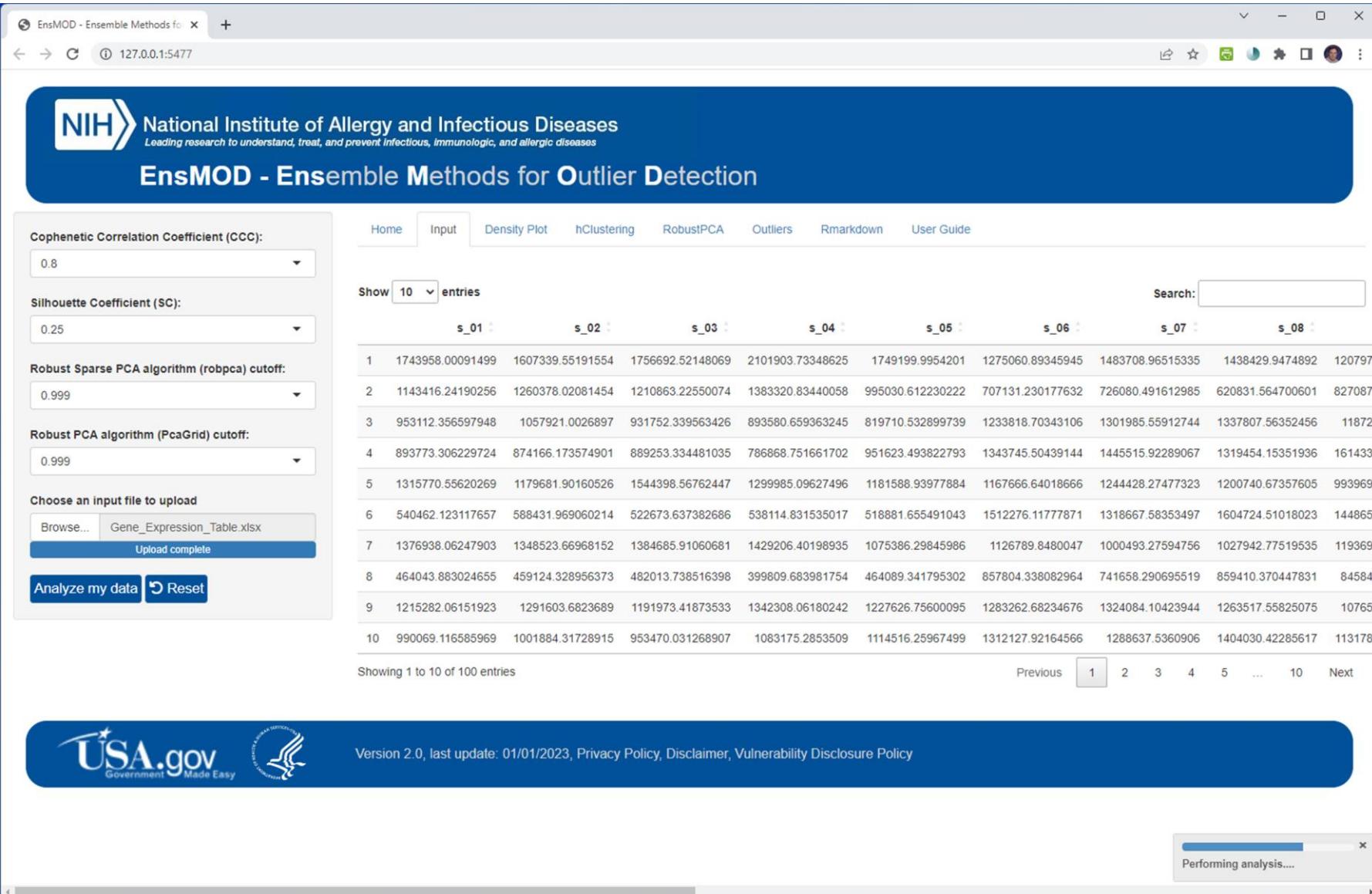
Previous 1 2 3 4 5 ... 10 Next

USA.gov Government Made Easy and **National Institute of Allergy and Infectious Diseases** logos.

Version 2.0, last update: 01/01/2023, Privacy Policy, Disclaimer, Vulnerability Disclosure Policy

Using EnsMOD

8. Click “Analyze my data”



The screenshot shows the EnsMOD - Ensemble Methods for Outlier Detection interface. On the left, there are several input parameters: Cophenetic Correlation Coefficient (CCC) set to 0.8, Silhouette Coefficient (SC) set to 0.25, Robust Sparse PCA algorithm (robPCA) cutoff set to 0.999, and Robust PCA algorithm (PcaGrid) cutoff set to 0.999. Below these is a file upload section where 'Gene_Expression_Table.xlsx' has been uploaded, indicated by a blue 'Upload complete' button. At the bottom left are 'Analyze my data' and 'Reset' buttons. The main area displays a table of 100 entries with columns labeled s_01 through s_08. A pink arrow points from the right towards the bottom right corner of the screen, which contains a progress bar with the text 'Performing analysis...'.

	s_01	s_02	s_03	s_04	s_05	s_06	s_07	s_08	
1	1743958.00091499	1607339.55191554	1756692.52148069	2101903.73348625	1749199.9954201	1275060.89345945	1483708.96515335	1438429.9474892	1207971
2	1143416.24190256	1260378.02081454	1210863.22550074	1383320.83440058	995030.612230222	707131.230177632	726080.491612985	620831.564700601	8270871
3	953112.356597948	1057921.0026897	931752.339563426	893580.659363245	819710.532899739	1233818.70343106	1301985.55912744	1337807.56352456	118720
4	893773.306229724	874166.173574901	889253.334481035	786868.751661702	951623.493822793	1343745.50439144	1445515.92289067	1319454.15351936	1614338
5	1315770.55620269	1179681.90160526	1544398.56762447	1299985.09627496	1181588.93977884	1167666.64018666	1244428.27477323	1200740.67357605	993969
6	540462.123117657	588431.969060214	522673.637382686	538114.831535017	518881.655491043	1512276.11777871	1318667.58353497	1604724.51018023	1448652
7	1376938.06247903	1348523.66968152	1384685.91060681	1429206.40198935	1075386.29845986	1126789.8480047	1000493.27594756	1027942.77519535	1193690
8	464043.883024655	459124.328956373	482013.738516398	399809.683981754	464089.341795302	857804.338082964	741658.290695519	859410.370447831	845845
9	1215282.06151923	1291603.6823689	1191973.41873533	1342308.06180242	1227626.75600095	1283262.68234676	1324084.10423944	1263517.55825075	107654
10	990069.116585969	1001884.31728915	953470.031268907	1083175.2853509	1114516.25967499	1312127.92164566	1288637.5360906	1404030.42285617	1131784



Using EnsMOD



EnsMOD HTML Output File of all Results is in \app\www\EnsMODoutputs\

The screenshot shows a web browser window with the title "Ensemble_Methods_for_Outlier_Detection_v2_0.html". The page content is as follows:

Ensemble_Methods_for_Outlier_Detection_Version_2.0

16 May, 2023

- 1. Set the Statistical Parameters and Prepare and Visualize the Input Data
- 2. Hierarchical Clustering for the Cophenetic Correlation Coefficient
- 3. Hierarchical Clustering for the Silhouette Coefficients
- 4. Robust PCA
 - 4.1 Robust Sparse PCA algorithm (robPCA)
 - 4.2 Robust PCA based on Projection Pursuit (PP) using GRID (PcaGrid)
- 5. Summary of the Results and Identification of Outlier(s)
 - 5.1 Outlier(s) Identified by Robust PCA analyses
 - 5.2 Outlier(s) Identified by Hierarchical Clustering
 - 5.3 Outlier(s) Identified by Both HCA and Robust PCA
- 6. References

1. Set the Statistical Parameters and Prepare and Visualize the Input Data

The four statistical cutoff values are set below. Your input file is loaded in here. The data are filtered by removing any row with "NaN" (missing values). Density plots are used to visualize the distribution of gene/protein expression values within each sample and across the samples. The normality of the dataset is visualized and tested. Note that it is unclear how accurately outliers will be detected if the input dataset is significantly different from a normal distribution.

```
# Set the Cophenetic Correlation Coefficient (CCC) cutoff (higher is more stringent).
# The CCC is a measure of how faithfully a dendrogram preserves the pairwise distances
# between the original unmodeled data points.
CCC_min <- as.numeric(params$CCC_min)
CCC_min
```

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

```
# Set the Silhouette Coefficient (SC) cutoff (Lower is more stringent).
# The SC is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation).
SC_max <- as.numeric(params$SC_max)
SC_max
```

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

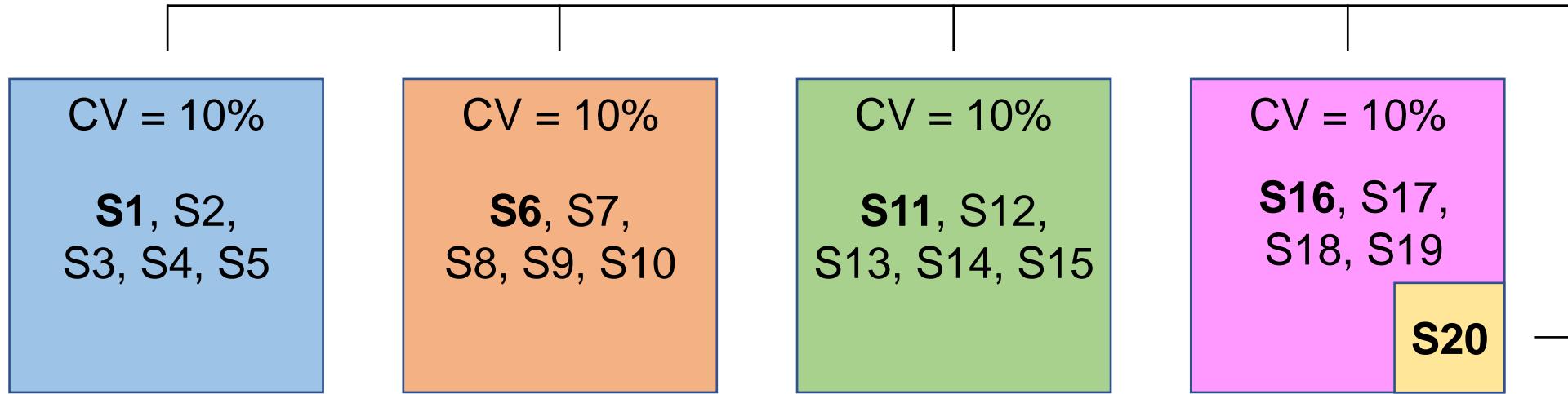
```
# Set the Robust PCA algorithm (robPCA) cutoff (higher is more stringent).
# For normally distributed data, this value is the estimated fraction of the samples that are not falsely classified as outliers.
robPCA_prob <- as.numeric(params$robPCA_prob)
robPCA_prob
```

```
## [1] 0.999
```

Simulated Proteomics Dataset

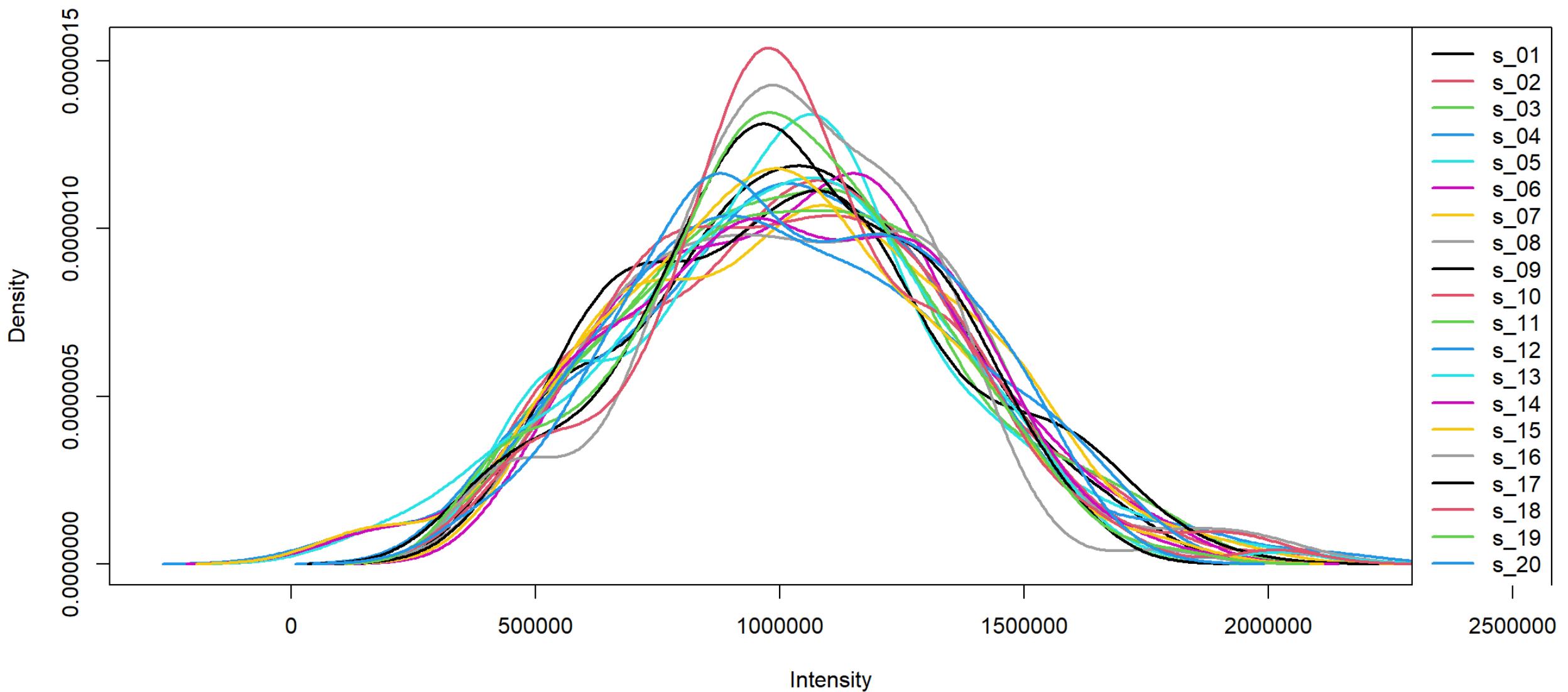
Simulated Proteomics Dataset

Coefficient of Variation (CV) = 30%

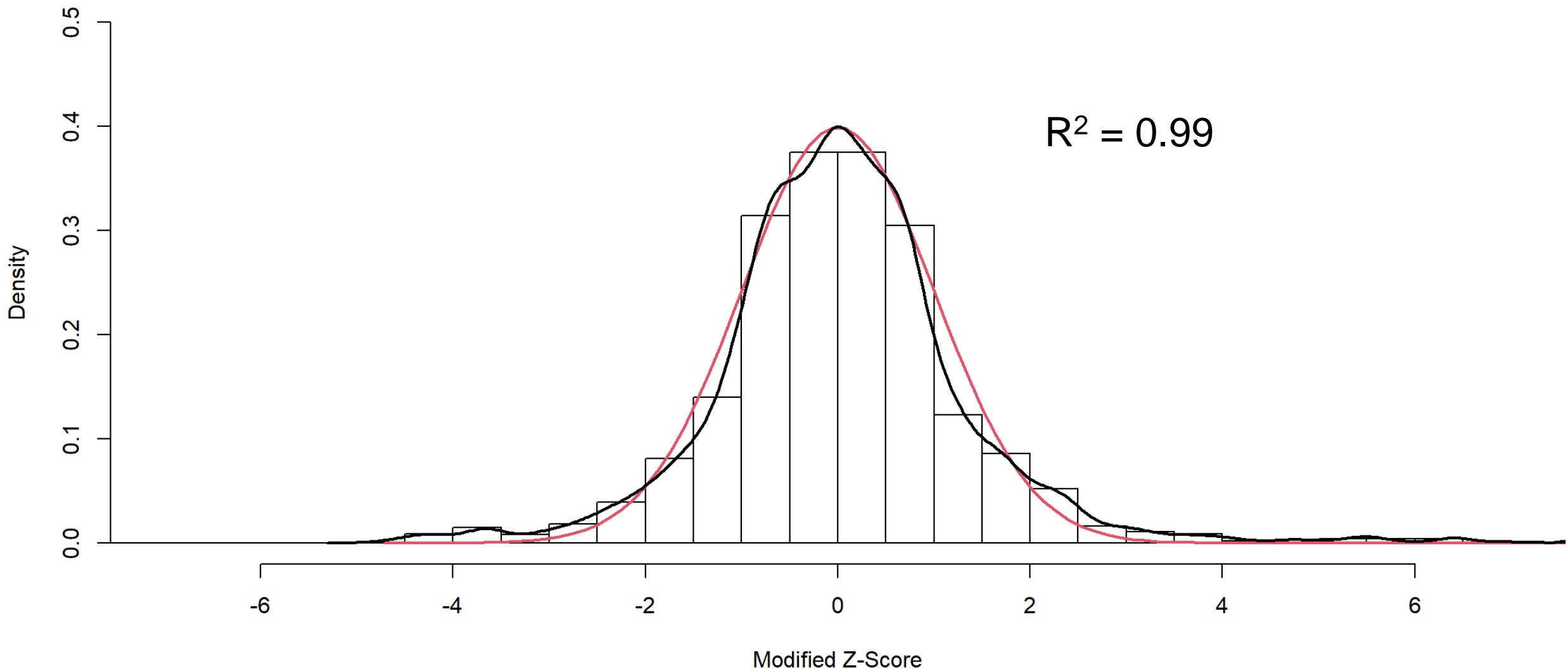


S1, S6, S11, S16, S20 = Initial Gaussian random sampling
 $\mu = 1,000,000$, CV = 30%, N=100 proteins

Densities of the Abundance Values for each Sample

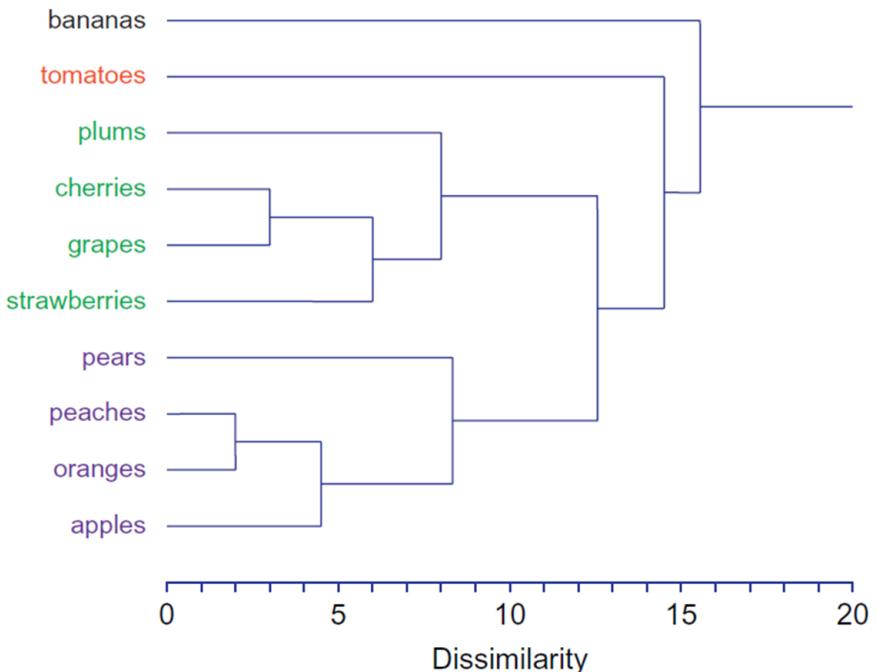


Empirical Histogram and Density versus the Standard Normal Distribution



Introduction to Hierarchical Cluster Analysis (HCA)

	Apples	Oranges	Strawberries	Bananas	Peaches	Plums	Tomatoes	Pears	Grapes	Cherries
Apples	—	5	11	16	4	10	12	8	11	10
Oranges		—	17	14	2	12	15	11	12	14
Strawberries			—	17	16	8	18	15	4	8
Bananas				—	17	15	20	11	14	16
Peaches					—	9	11	6	15	13
Plums						—	12	10	9	7
Tomatoes							—	16	18	14
Pears								—	12	14
Grapes									—	3
Cherries										—



Hierarchical Cluster Analyses and the Cophenetic Correlation Coefficient (CCC)

3 Distances: Euclidean, Manhattan, Pearson Correlation

5 Linkages: Single, Average, Complete, Centroid, Ward

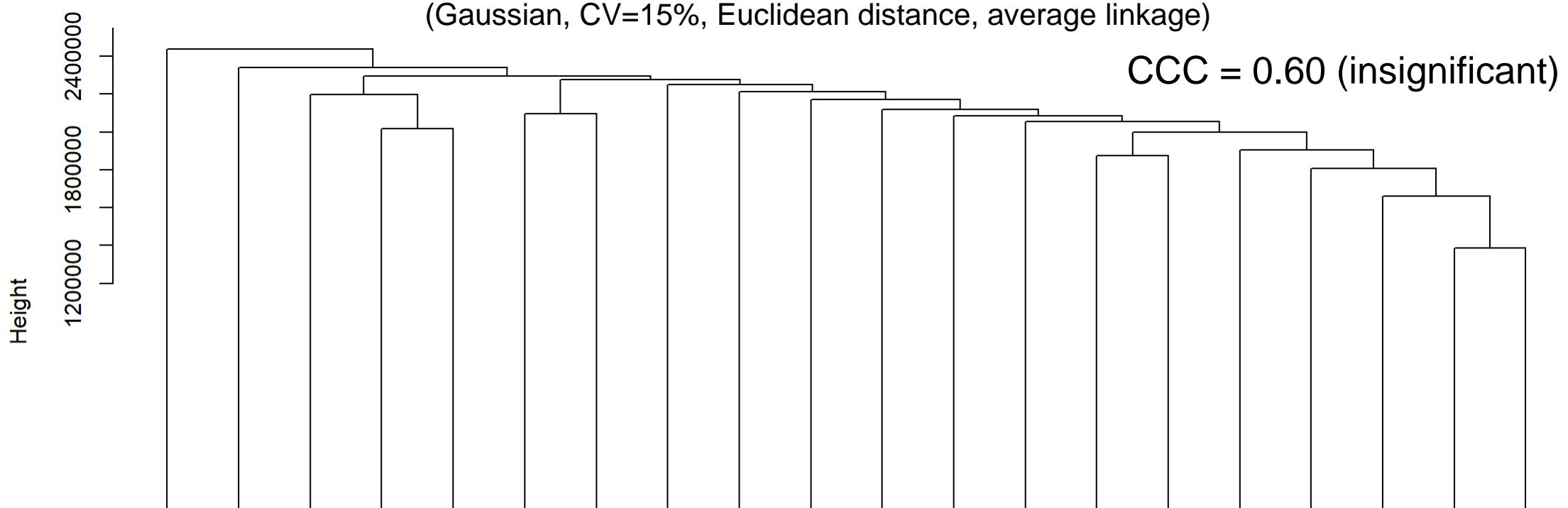
$3 \times 5 = 15$ HCAs total

The HCA with the largest CCC is selected for downstream analyses

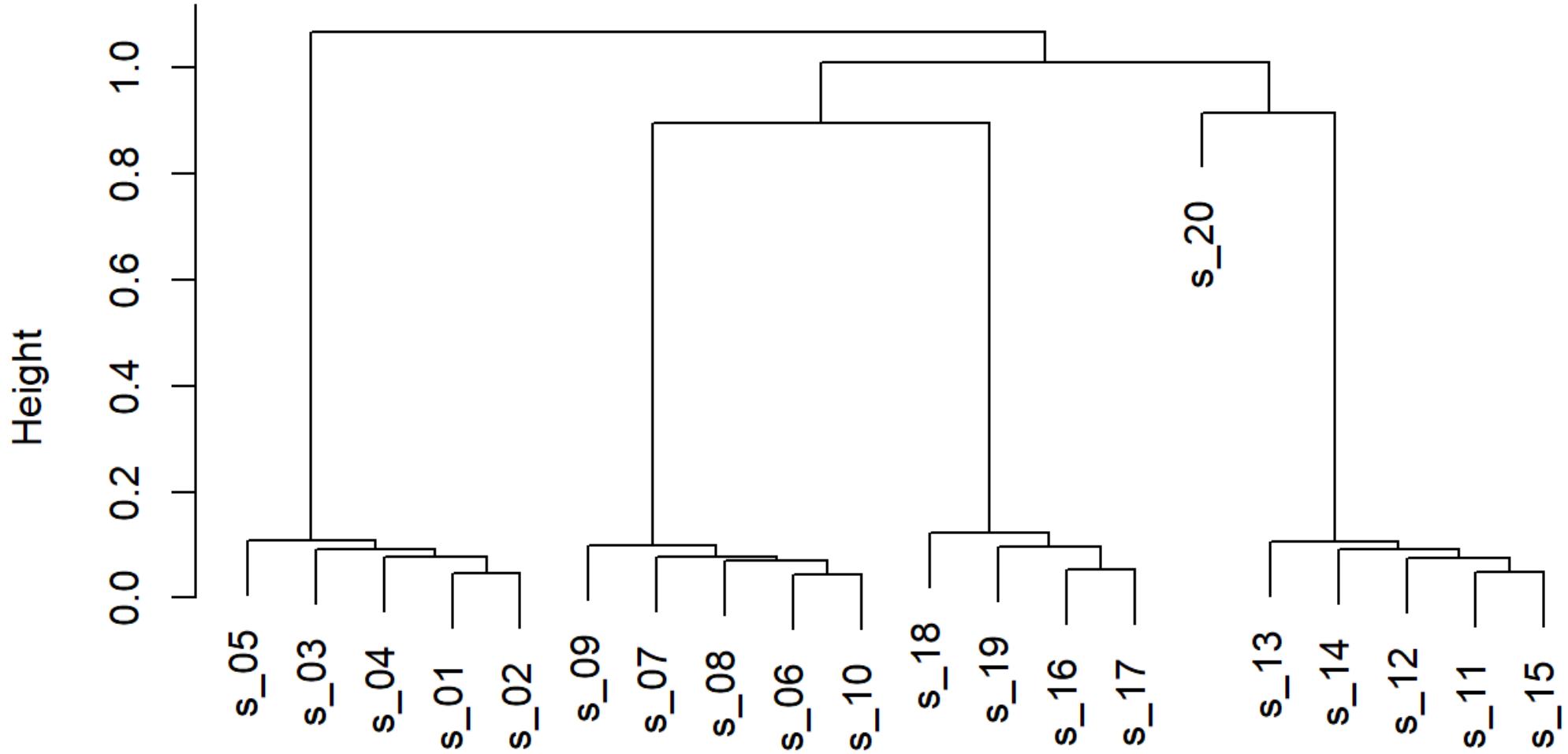
If $\text{CCC} \geq 0.8$, then outliers might be identifiable [Selicato et al 2021 Mathematics 9:882]

Example of poor clustering: HCA of purely random data

(Gaussian, CV=15%, Euclidean distance, average linkage)



If CCC ≥ 0.8 (default parameter value), then outliers might be identifiable



CCC = 0.99 (Pearson corr. distance, avg. linkage)

Silhouette Coefficient (SC)

The SC of a sample within an HCA quantifies:

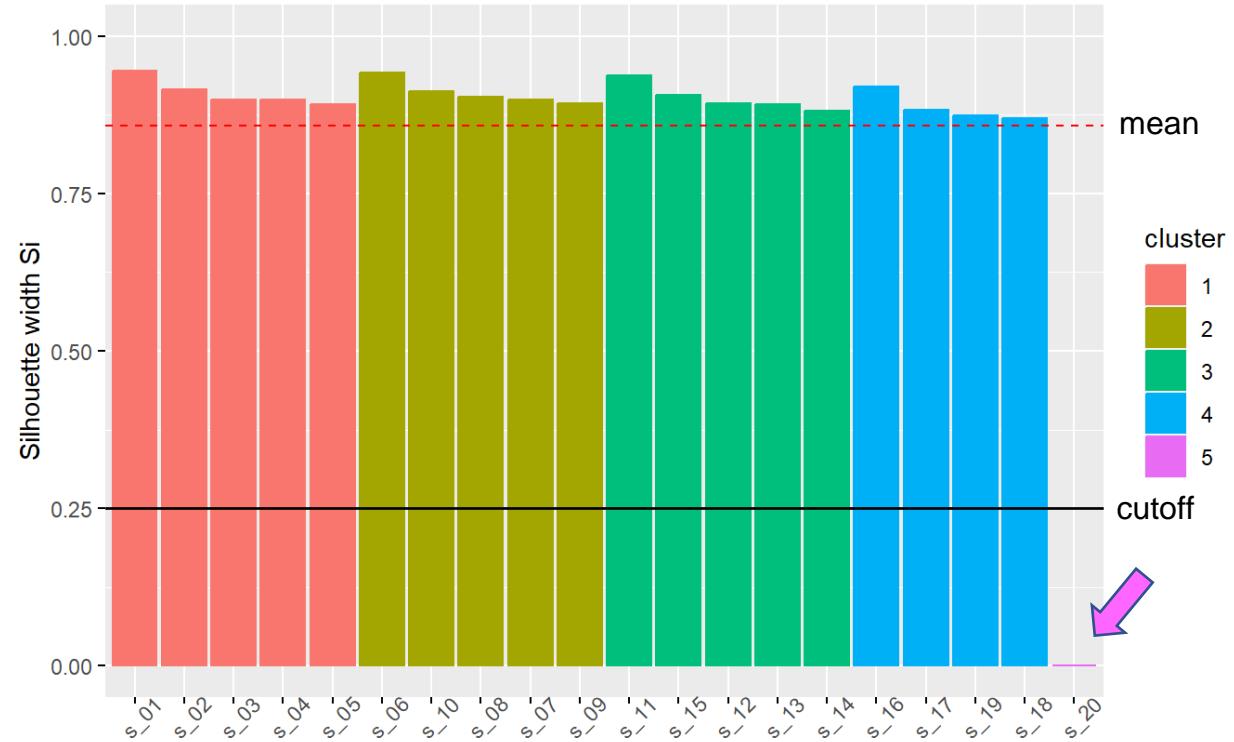
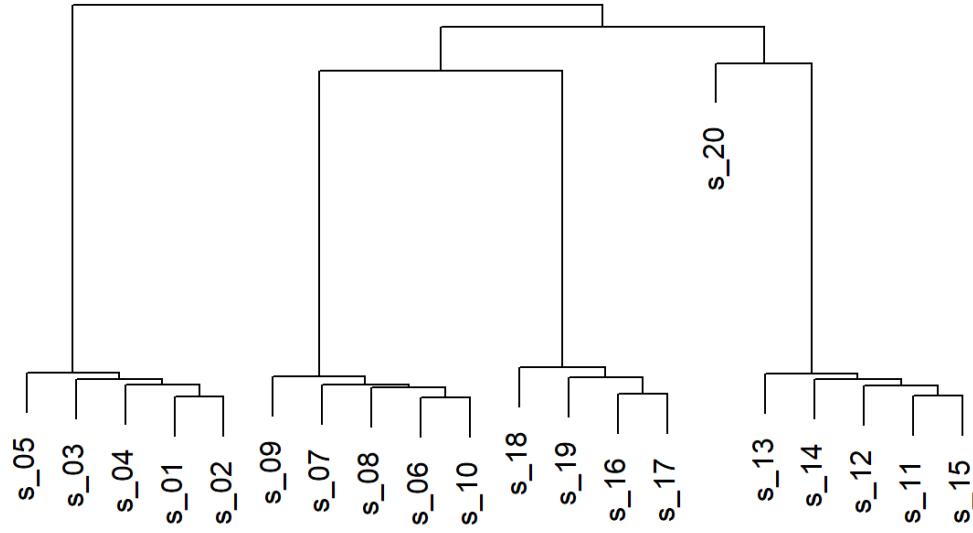
How well the sample fits within its own HCA cluster
versus

How well the sample would fit in the other HCA clusters

Table 1. Structures corresponding to particular Silhouette coefficient ranges.

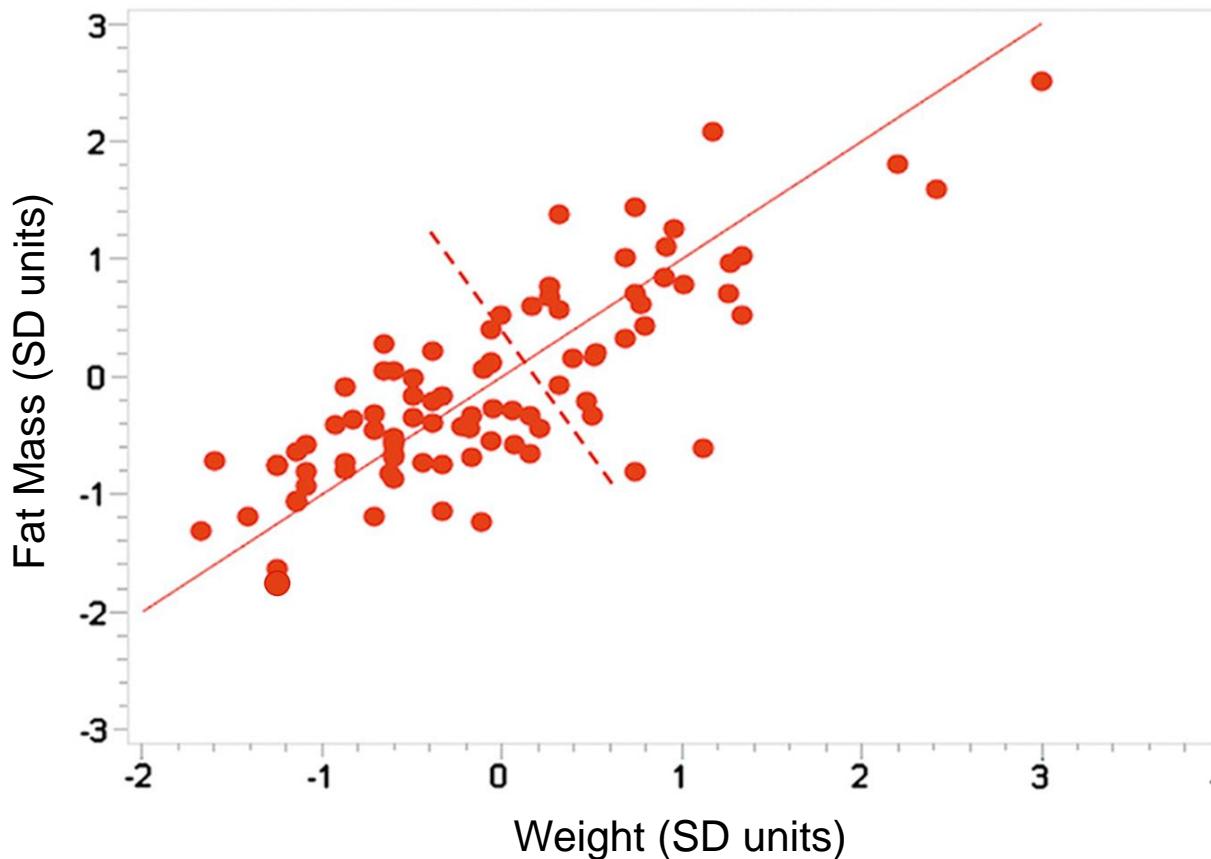
SC	Interpretation
0.71–1.0	A strong structure was found.
0.51–0.70	A reasonable structure was found.
0.26–0.50	The structure is weak and may be artificial.
<0.25	No substantial structures have been found.

If $SC < 0.25$ (default parameter value), then the sample might be an outlier

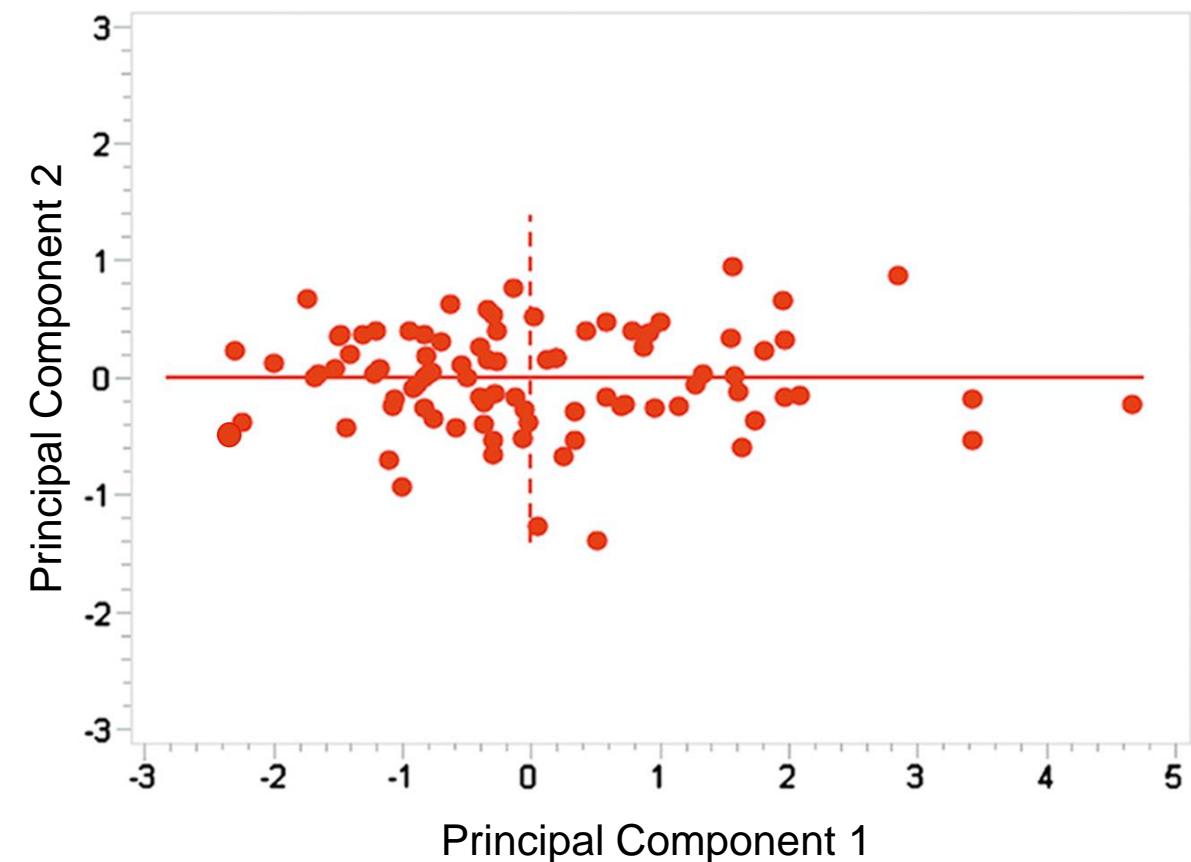


$SC[s_20] = 0$

Introduction to Principal Component Analysis (PCA)



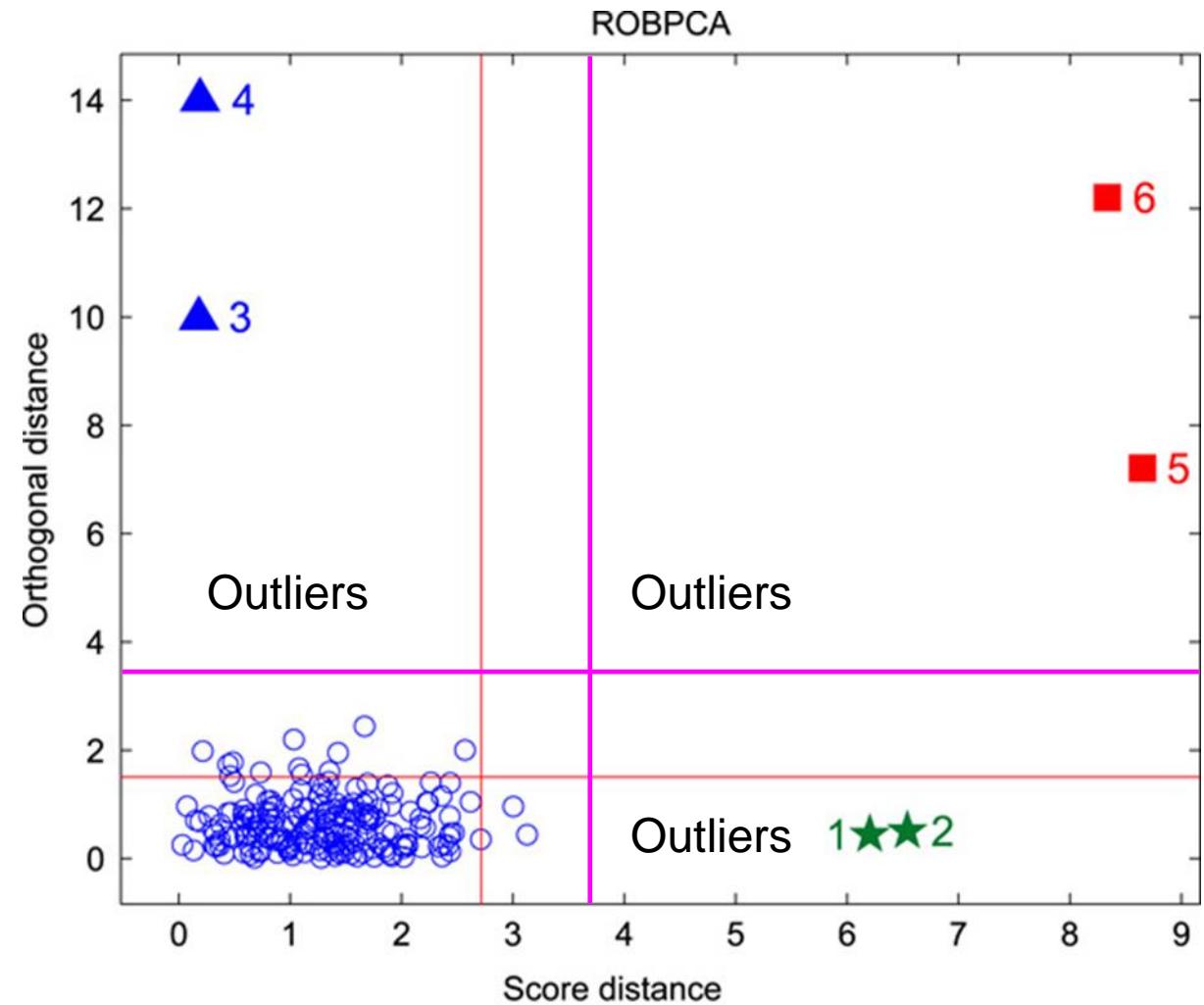
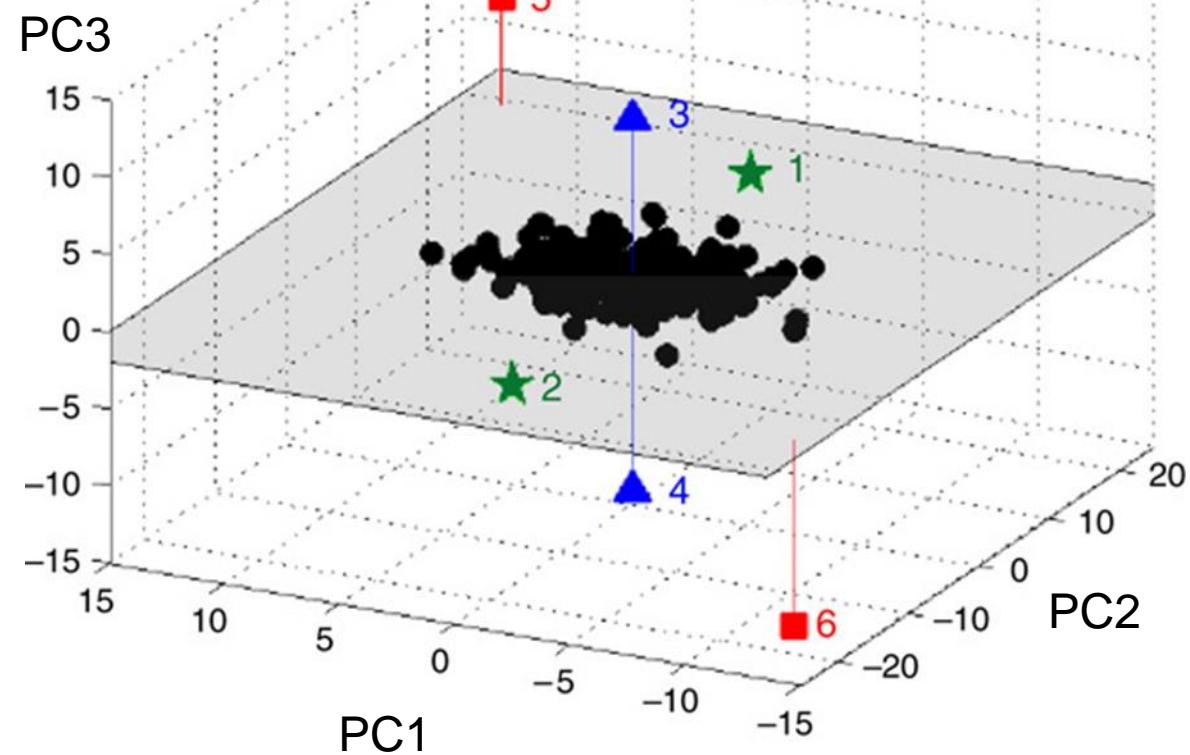
SD = standard deviation



PCA: mean center and rotate for maximum variance along PC1, then PC2, ..., then PCn

Robust PCA: a PCA algorithm that is insensitive to extreme points

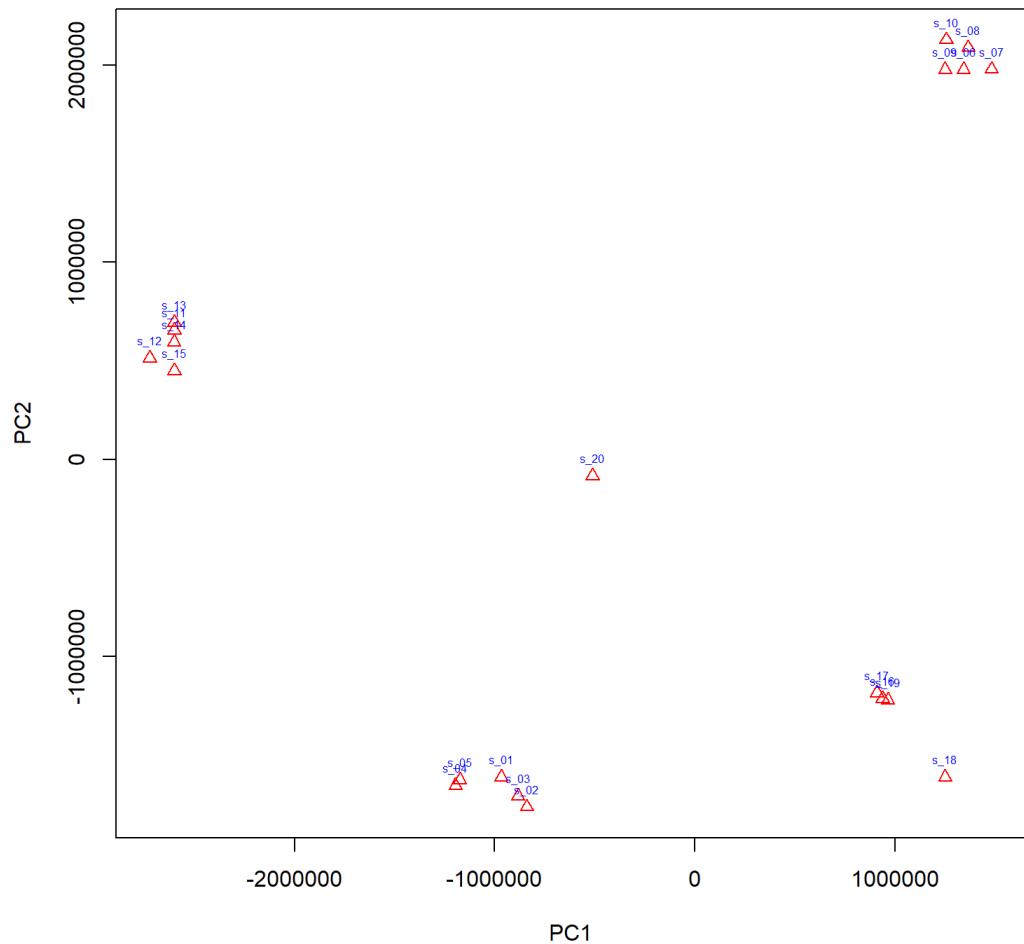
Example of a PCA Distance-Distance Plot



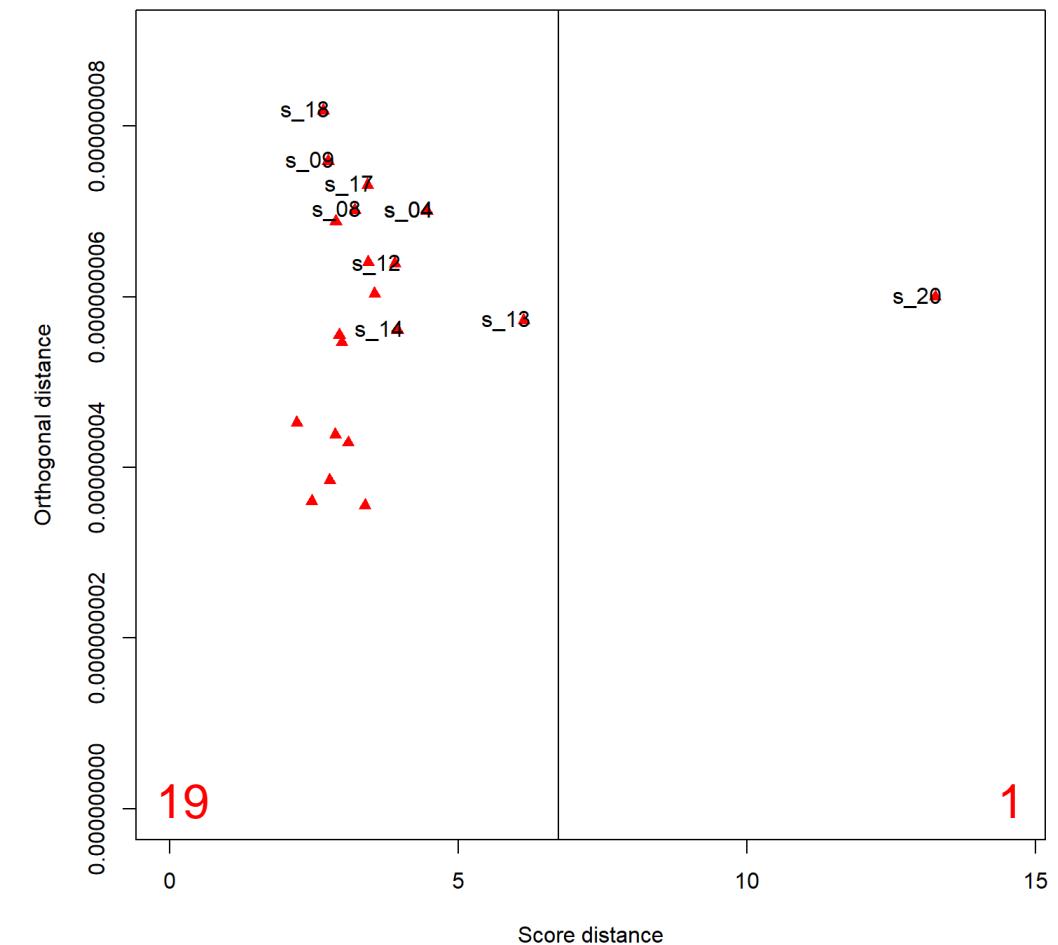
rPCA parameter: 97.5%. Est. Confidence = $(97.5\%)^2 = 95.1\%$
rPCA parameter: 99.9%. Est. Confidence = $(99.9\%)^2 = 99.8\%$

Two Robust Principal Component Analyses are Performed

Robust PCA (PcaGrid) BiPlot



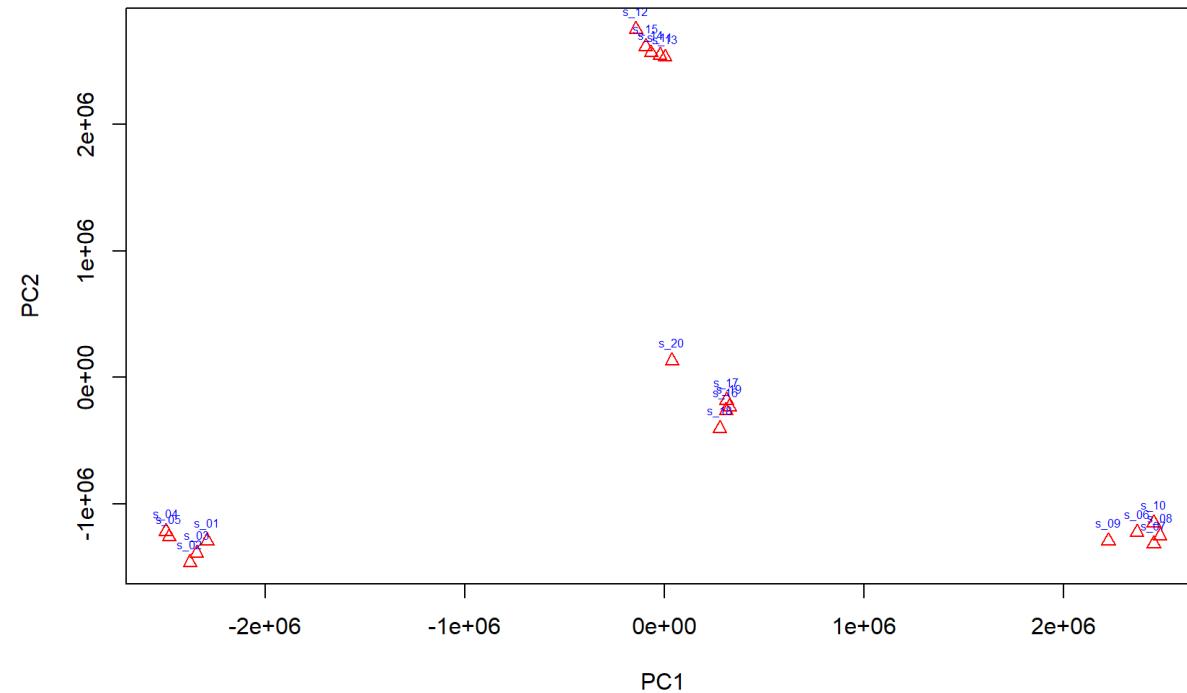
Robust PCA (PcaGrid)



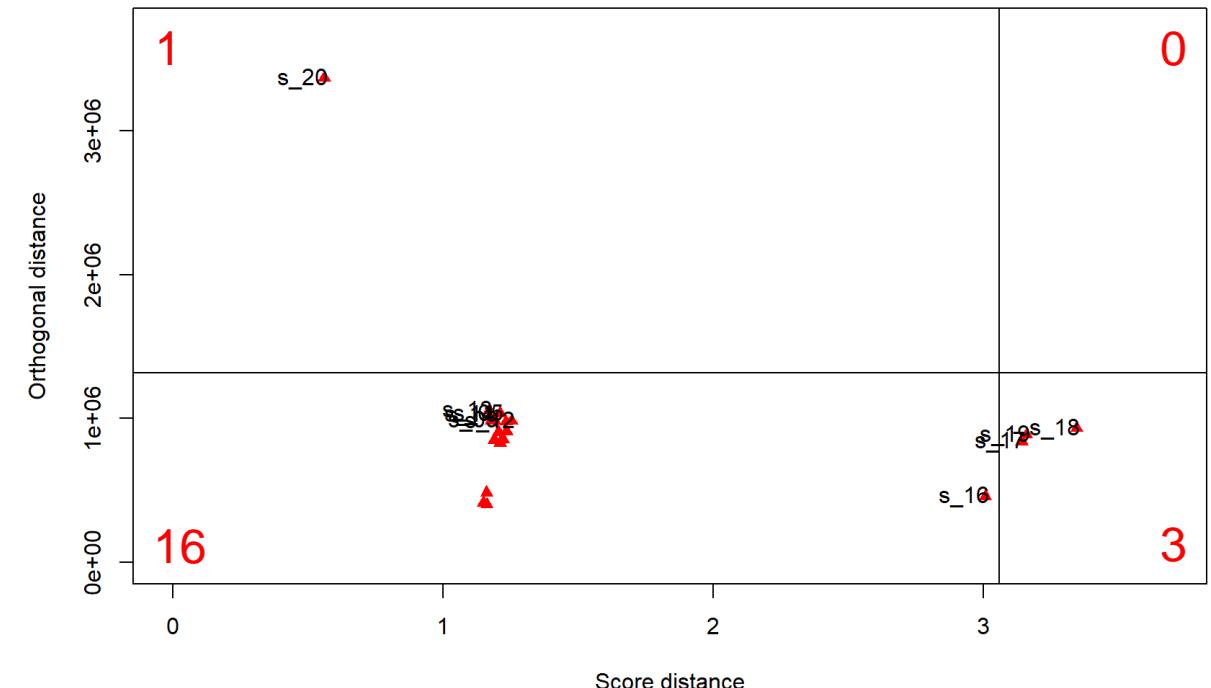
rPCA parameter was set to 99.9%
Estimated Confidence = $(99.9\%)^2 = 99.8\%$

Two Robust Principal Component Analyses are Performed

Robust PCA (robPCA) BiPlot



Robust PCA (robPCA)



rPCA parameter was set to 99.9%
Estimated Confidence = $(99.9\%)^2 = 99.8\%$

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"
```

RESEARCH ARTICLE

Open Access



Check for
updates

Robust principal component analysis for accurate outlier sample detection in RNA-Seq data

Xiaoying Chen¹, Bo Zhang², Ting Wang^{3,4}, Azad Bonni¹ and Guoyan Zhao^{1*}

* Correspondence: gzhao@wustl.edu

¹Department of Neuroscience,
Washington University School of
Medicine, St. Louis, MO, USA

Full
avail

Abstract

Background: High throughput RNA sequencing is a powerful approach to study gene expression. Due to the complex multiple-steps protocols in data acquisition,

PcaGrid
 $p = 97.5\%$
Recommend: $p = 99.9\%$

and used the result as reference to compare the performance of eight different data

Article

A New Ensemble Method for Detecting Anomalies in Gene Expression Matrices

Laura Selicato ^{1,2,*†} , Flavia Esposito ^{1,2,†} , Grazia Gargano ¹ , Maria Carmela Vegliante ³ , Giuseppina Opinto ³ , Gian Maria Zaccaria ³ , Sabino Ciavarella ³ , Attilio Guarini ³ and Nicoletta Del Buono ^{1,2}

<https://doi.org/10.3390/math8030369> (F.E.);

HCA CCC ≥ 0.8

HCA SC < 0.25

ch up
Citation:
Gargano,
G.; Zacc
Guarini,
Ensembl
Anomaly
Matrices.
<https://doi.org/10.3390/math8030369>
Academic Editor: Junseok Kim

robPCA
 $p = 97.5\%$
Recommend: $p = 99.9\%$

Real datasets often contain observations that behave differently from the majority of the data. If an occurrence differs from the dominant part of the data, or if it is sufficiently unlikely under the assumed data probability model, it is considered an anomaly or outlier.

Anthrax Phosphoproteomics

Spleen Phosphoproteomics of Mice with Anthrax

Toxin-,
capsule-,
asymptomatic,
abortive infection

Toxin+,
capsule-,
lethal
in 2-4d

Prepared: 5 mice 15 mice 15 mice 25 mice

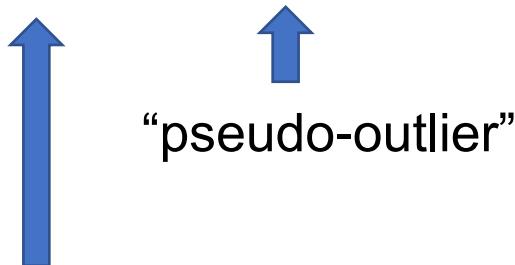
Time	No Injection	Vehicle	ΔSterne	Sterne
0 h	5 mice			
24 h		5 mice	5 mice	5 mice
48 h		5 mice	5 mice	5 mice
72 h		5 mice	5 mice	1 mouse

The Sterne 72h sample functioned as a pseudo-outlier

Spleens homogenized, trypsin, phosphopeptides enriched, LC-MS(/MS), Log10 label-free quantitation

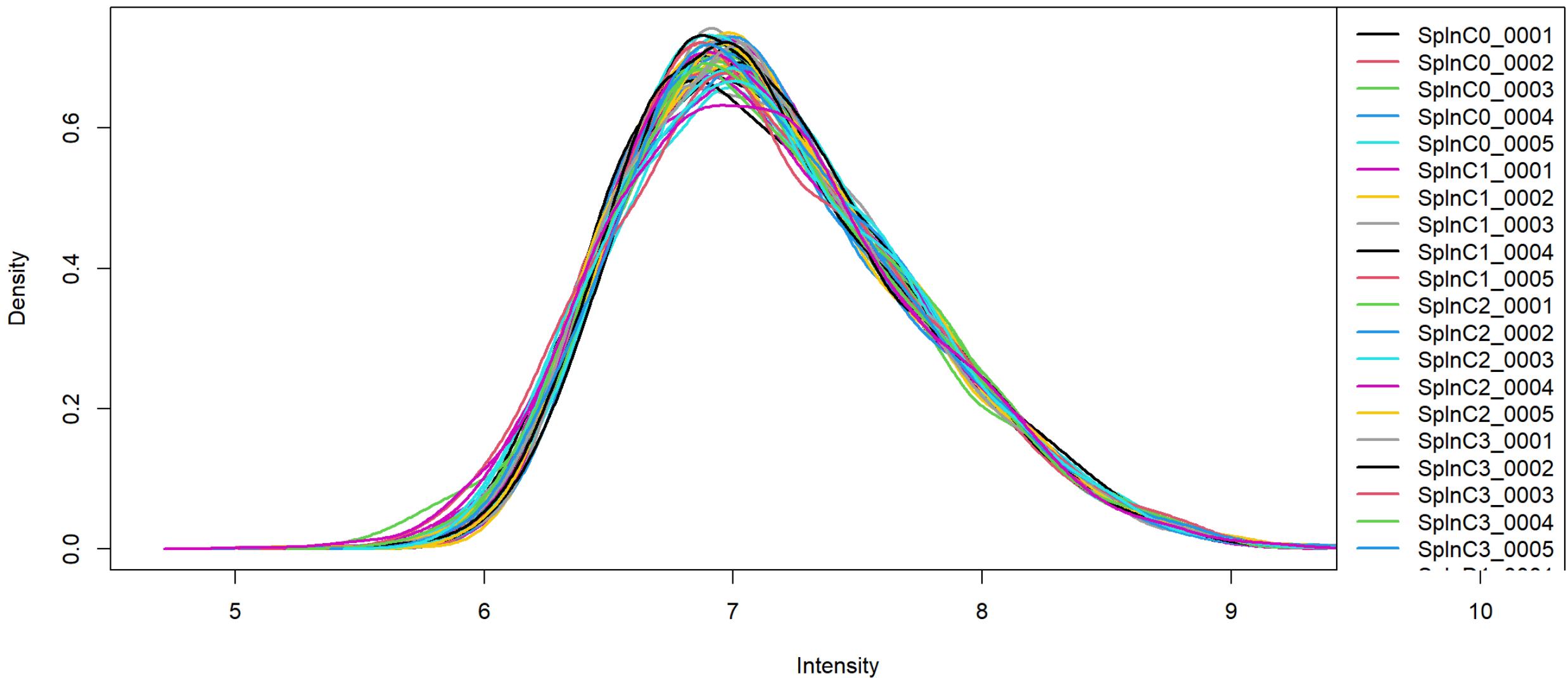
5.3 Outlier(s) Identified by Both HCA and Robust PCA

```
# 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] "SplnC0_0003" "SplnS3_0001"
```



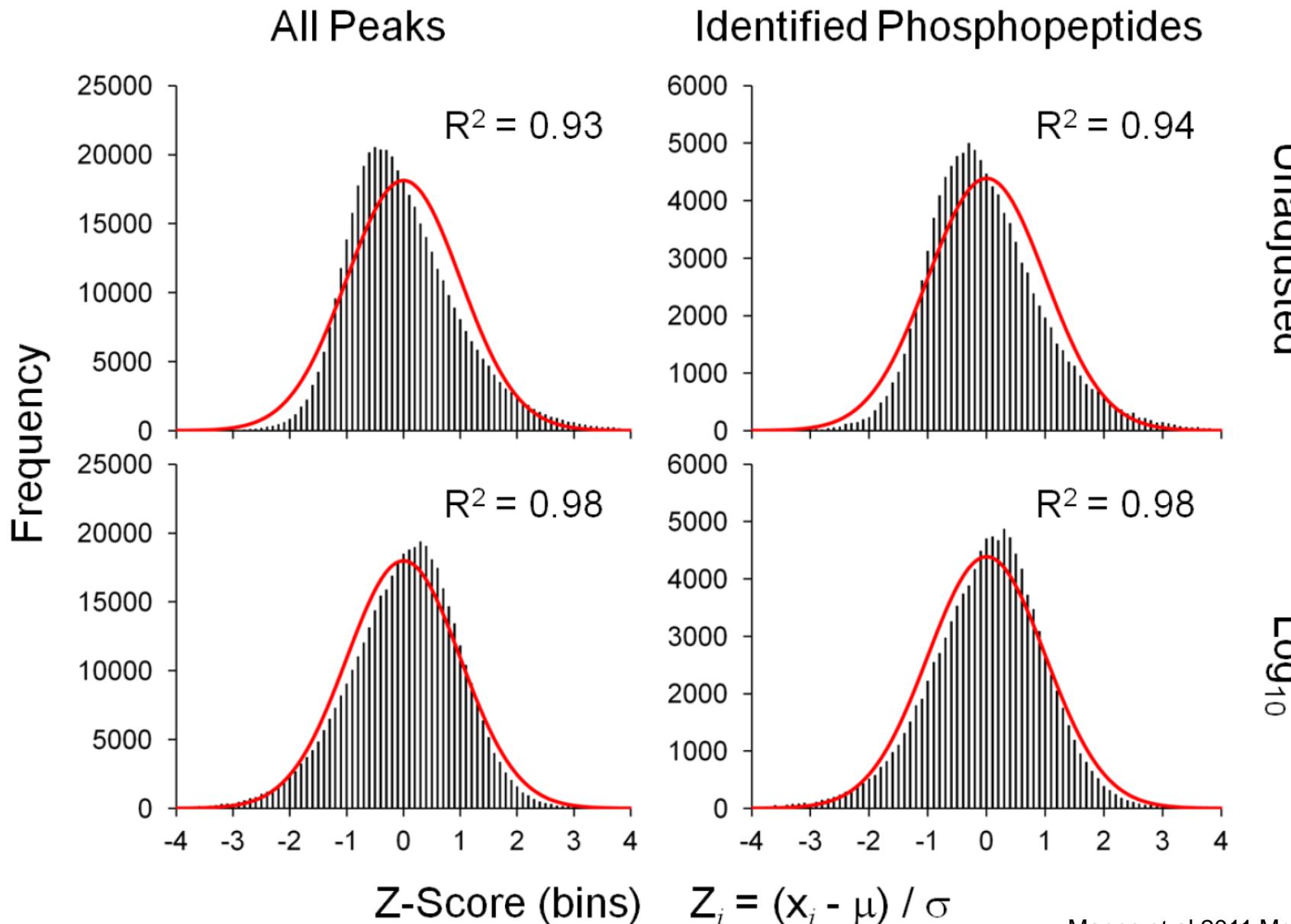
One of the 20 uninfected samples

Densities of the Abundance Values for each Sample

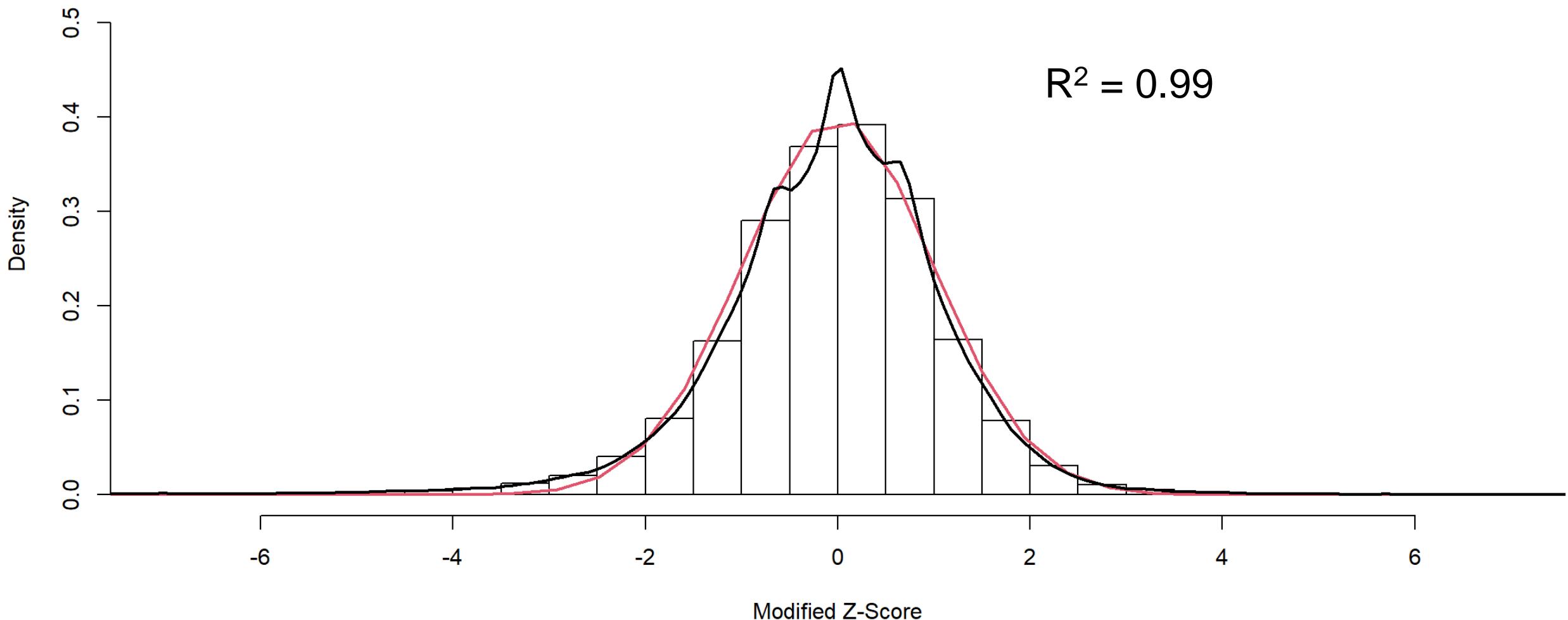


Example of Data Transformation to Increase Normality

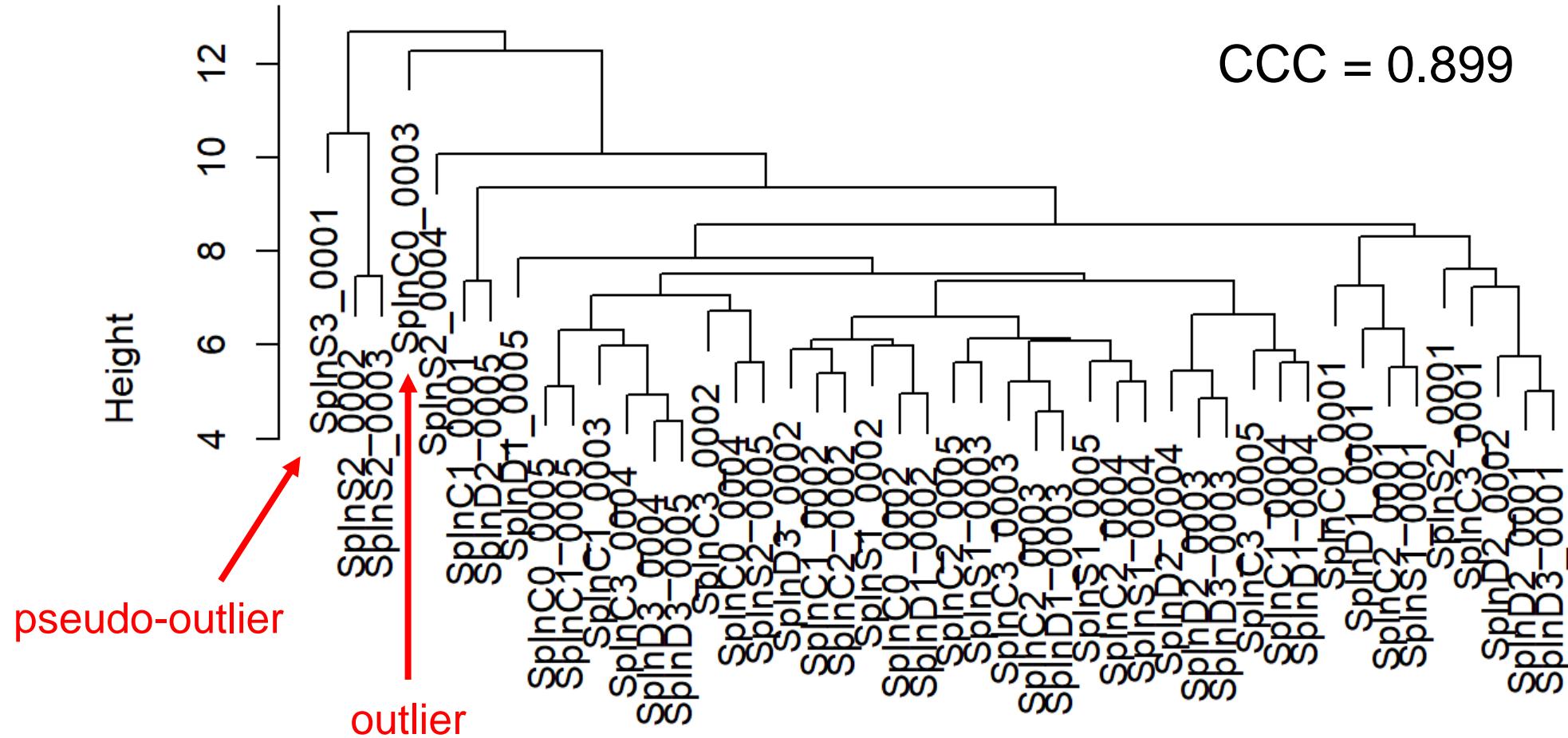
LC-MS Intensity Data are Log-Normal



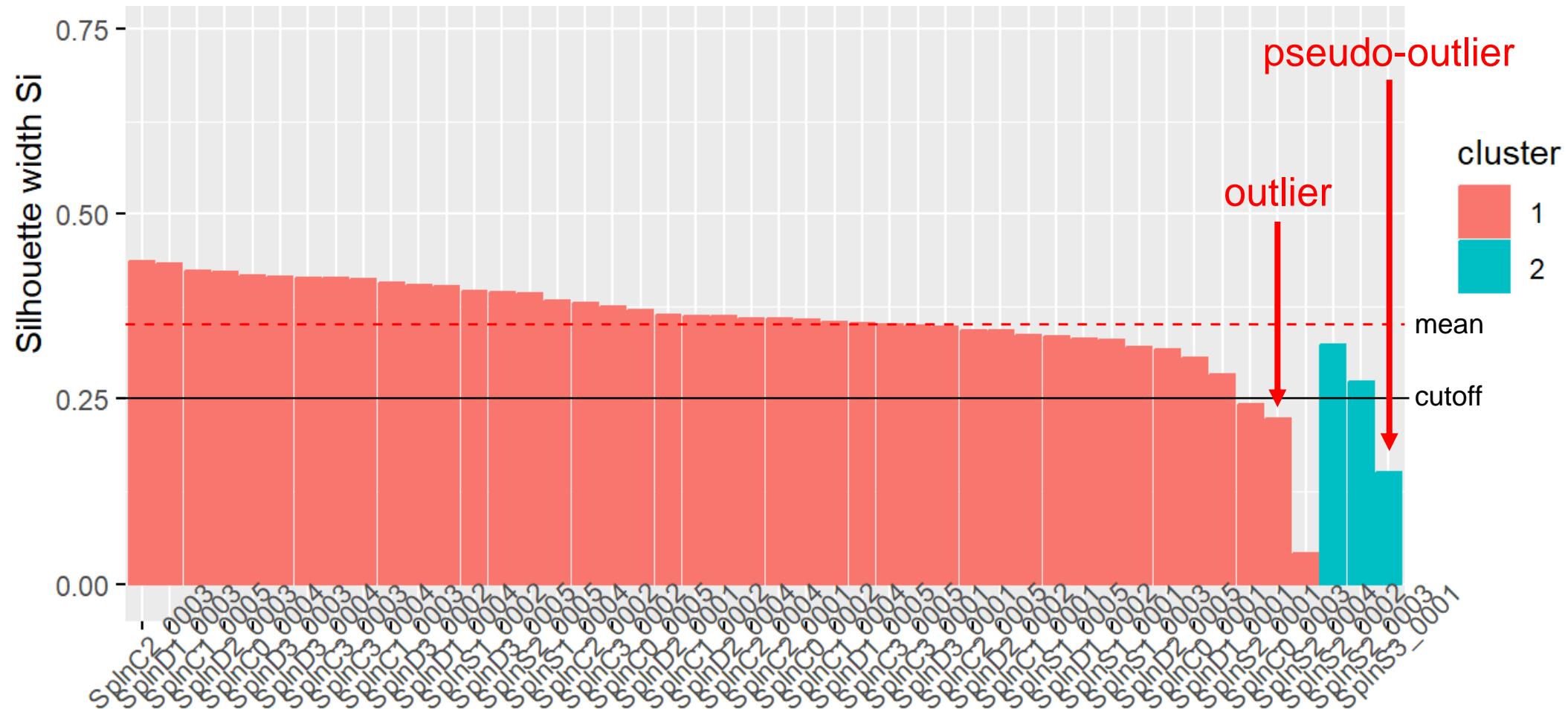
Empirical Histogram and Density versus the Standard Normal Distribution



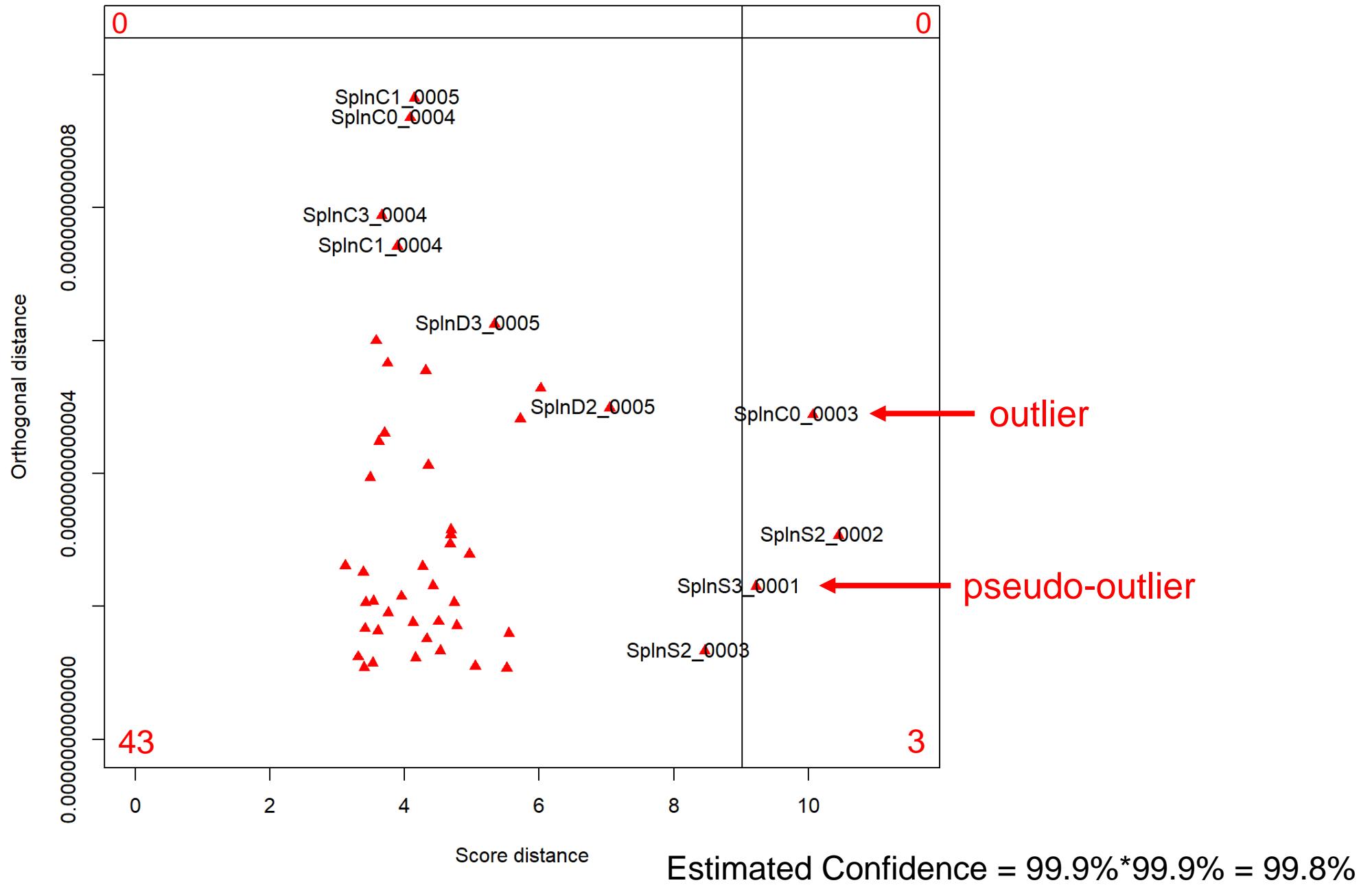
If CCC ≥ 0.8 (default parameter value), then outliers might be identifiable



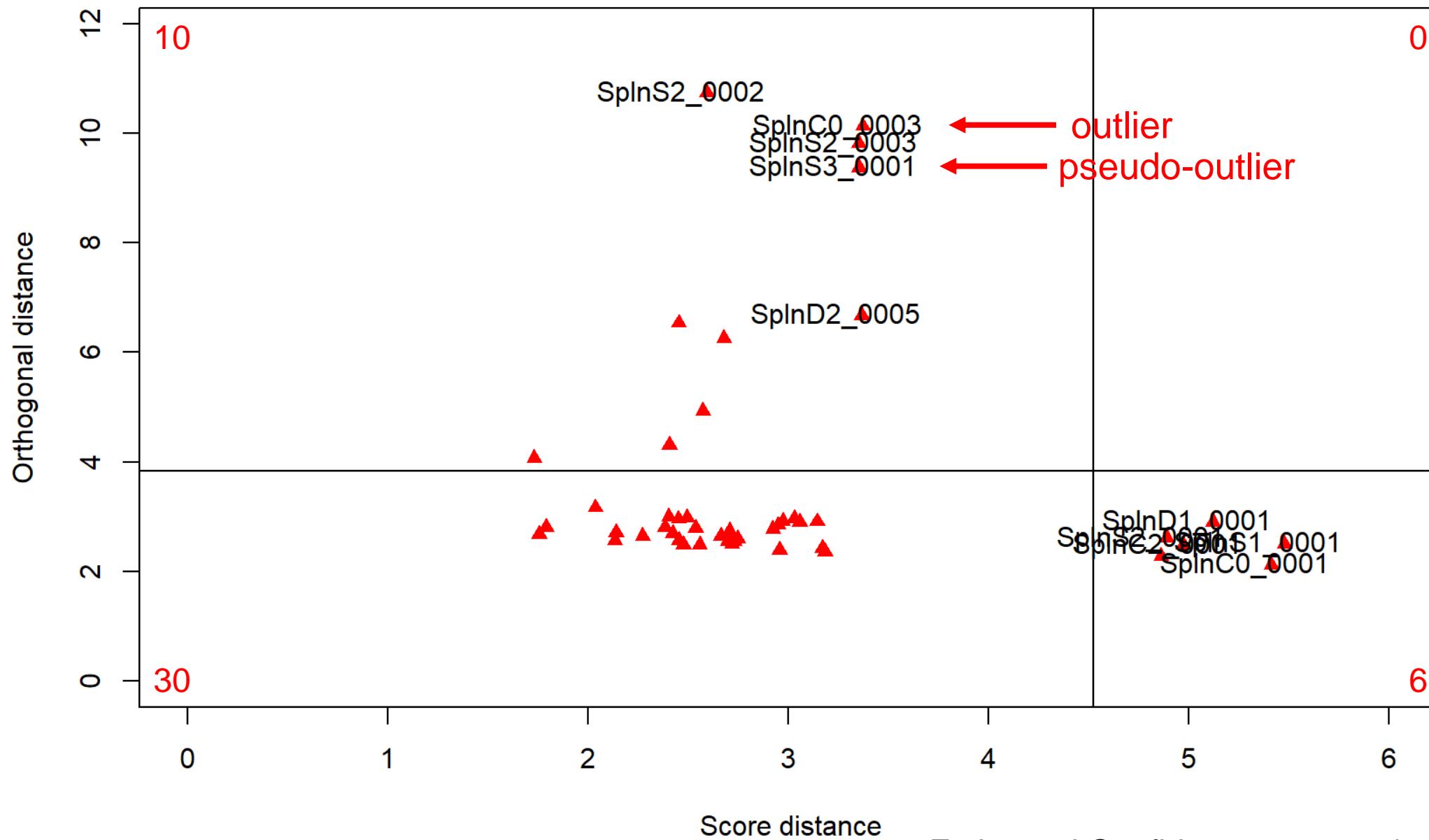
If $SC < 0.25$ (default parameter value), then the sample might be an outlier



Robust PCA (PcaGrid)

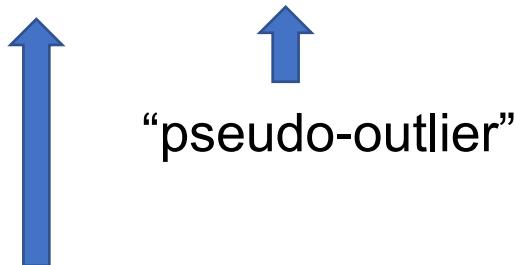


Robust PCA (robPCA)



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] "SplnC0_0003" "SplnS3_0001"
```



One of the 20 uninfected samples

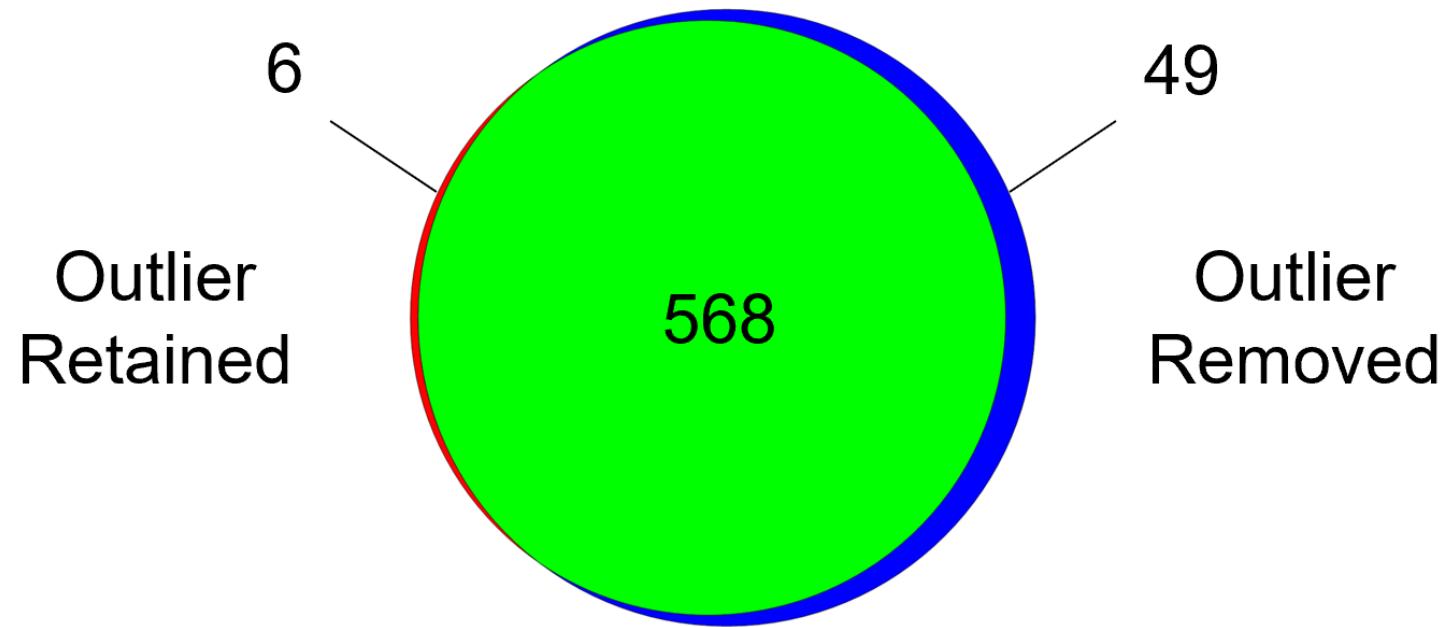
Could a single outlier among 20 negative control samples significantly impact downstream statistical analyses?

1-way ANOVAs were performed (q -value ≤ 0.05)

Both with and without including the true outlier SplnC0_0003

The pseudo-outlier SplnS3_0001 was not included in either set of ANOVA analyses

All of the uninfected mice were grouped into a single experimental condition



Detecting and removing the outlier apparently improved the accuracy of the results

Conclusions

- Outlier detection can help with:
 - Preventing erroneous biological conclusions
 - Improving experimental protocols
 - Discovering rare biological mechanisms
- EnsMOD incorporates two published algorithms for transcriptomics sample outlier detection
- EnsMOD can be used to analyze any omics dataset (~Gaussian variance, ≥ 9 samples)
- EnsMOD
 - Plots density curves
 - Measures data variance normality
 - Performs hierarchical cluster analyses
 - Performs robust principal component analyses
- EnsMOD successfully identified the simulated outliers
- Phosphoproteomics sample outlier detection and removal resulted in additional ANOVA hits
- EnsMOD is free, open-source, easy to use, and published