Ensemble Methods for Outlier Detection (EnsMOD):

A Software Program for Omics Sample Outlier Detection

June 1, 2023

NIH National Institute of Allergy and Infectious Diseases

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 30, Number 0, 2023

Pp. 1–10
DOI: 10.1089/cmb.2022.0243

Software

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

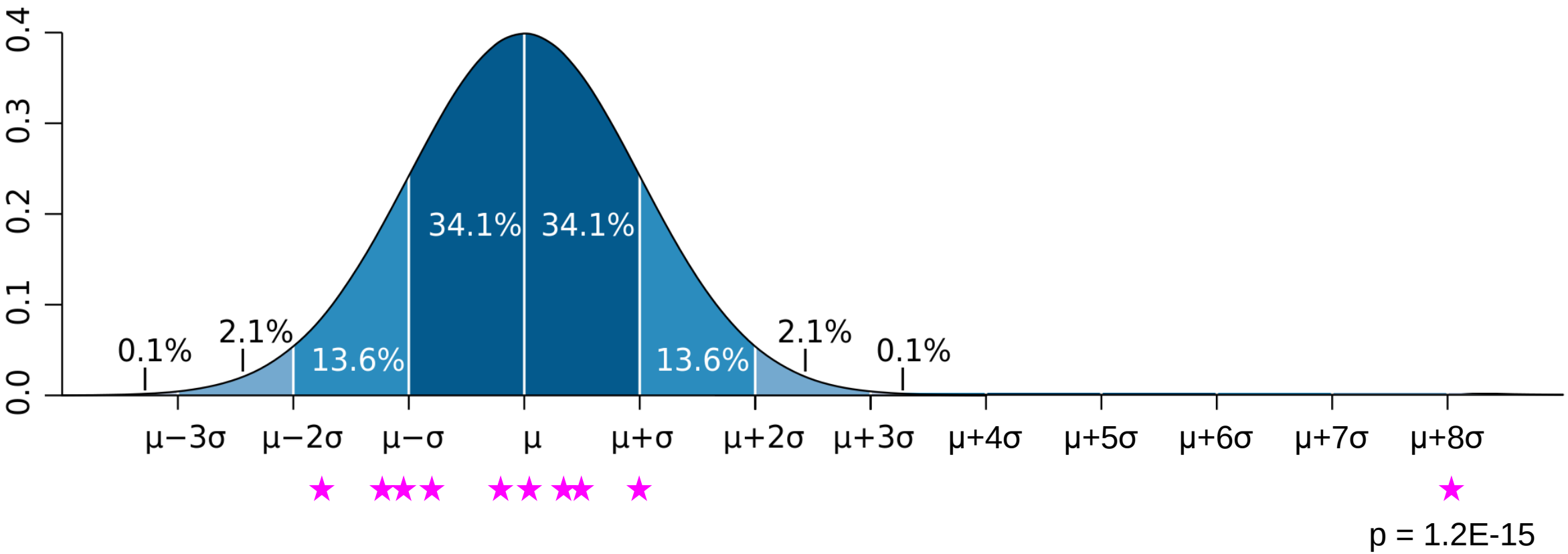
NATHAN P. MANES,* JIAN SONG,* and ALEKSANDRA NITA-LAZAR

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."



https://en.wikipedia.org/wiki/File:Standard_deviation_diagram_micro.svg

Hawkins DM. Identification of Outliers. Springer International Publishing AG, Cham, Switzerland; 1980; doi: 10.1007/978-94-015-3994-4

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

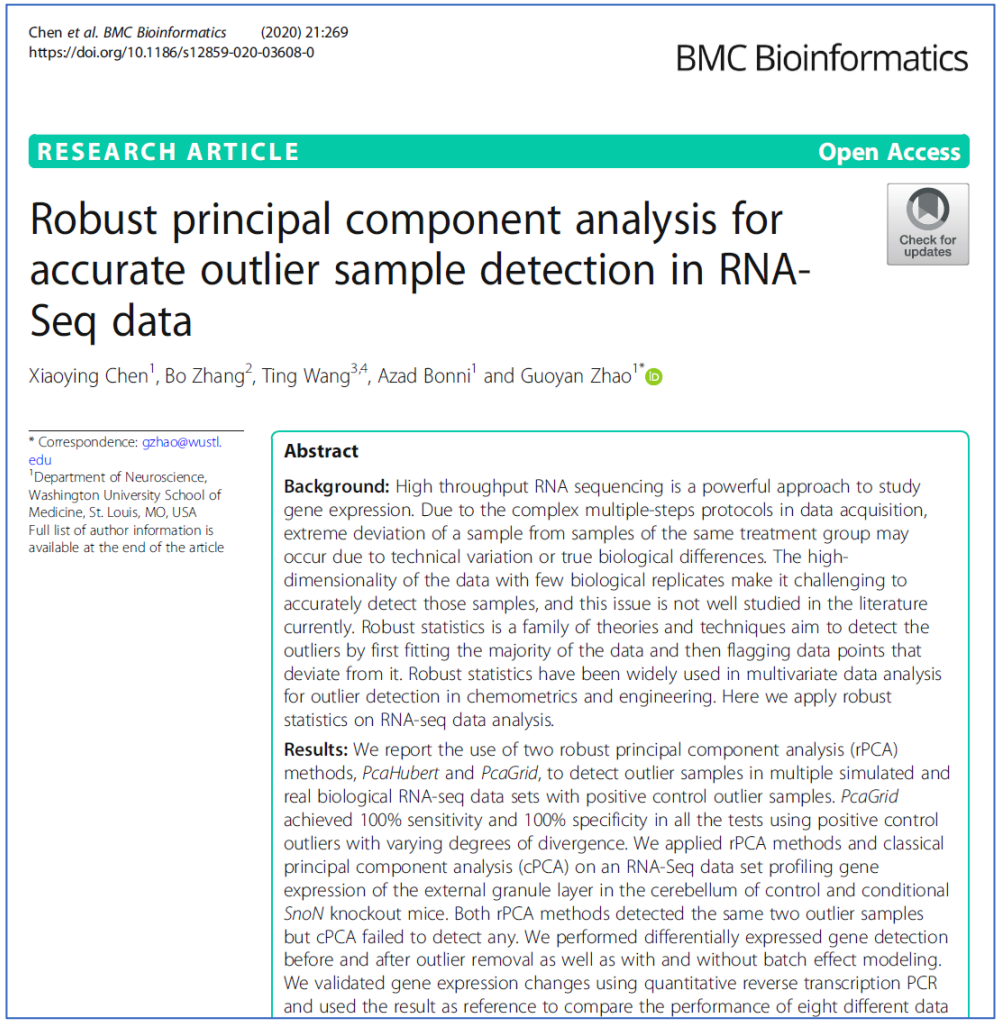
Example 1: LC-MS of a tissue proteomics sample, but the vial had an air bubble resulting in an unintended blank run

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

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

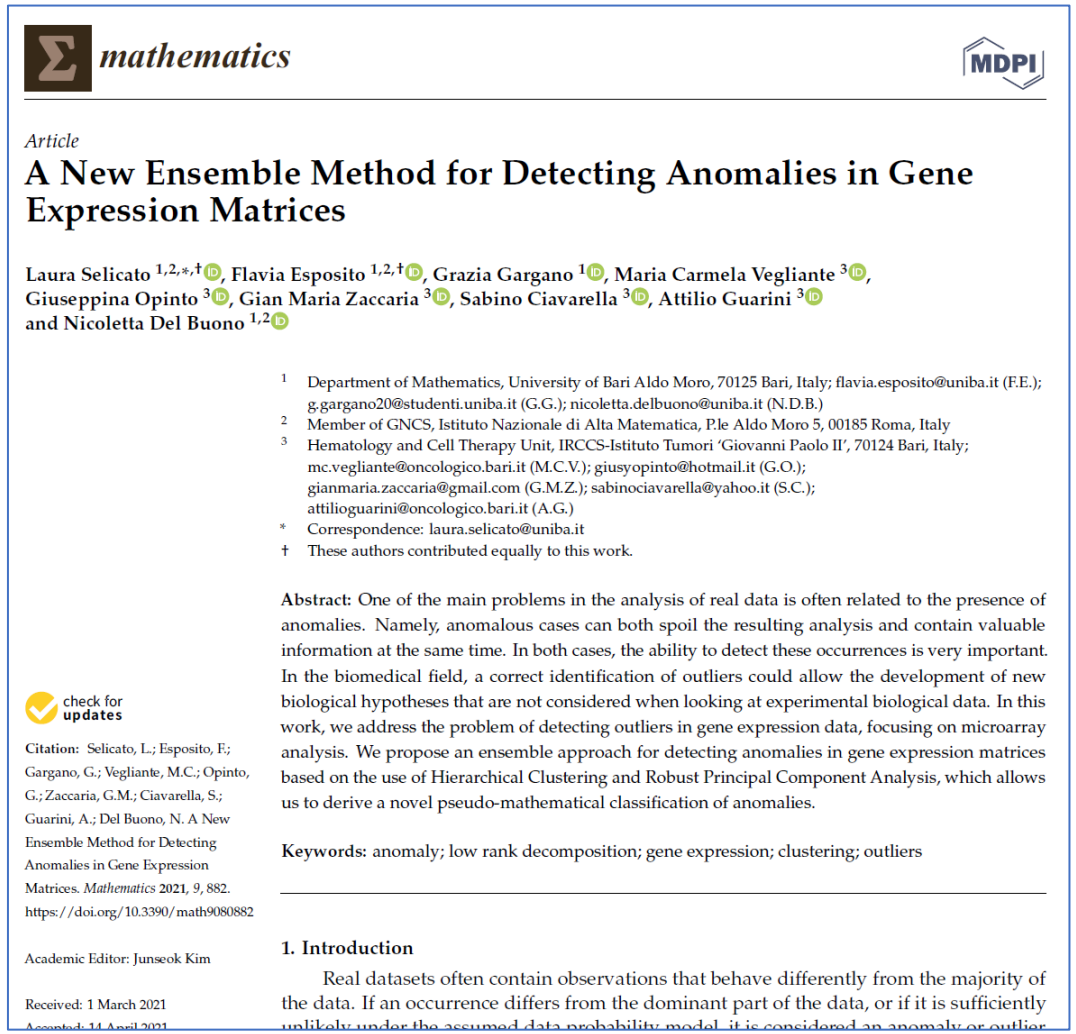
Xiaoying Chen[1], Bo Zhang[2], Ting Wang[3,4], Azad Bonni[1] and Guoyan Zhao[1*]

* Correspondence: gzhao@wustl.edu
[1]Department 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

MDPI

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, P.le 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.); gianmaria.zaccaria@gmail.com (G.M.Z.); sabinociavarella@yahoo.it (S.C.); attilioguarini@oncologico.bari.it (A.G.)
* Correspondence: laura.selicato@uniba.it
† These authors contributed equally to this work.

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

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

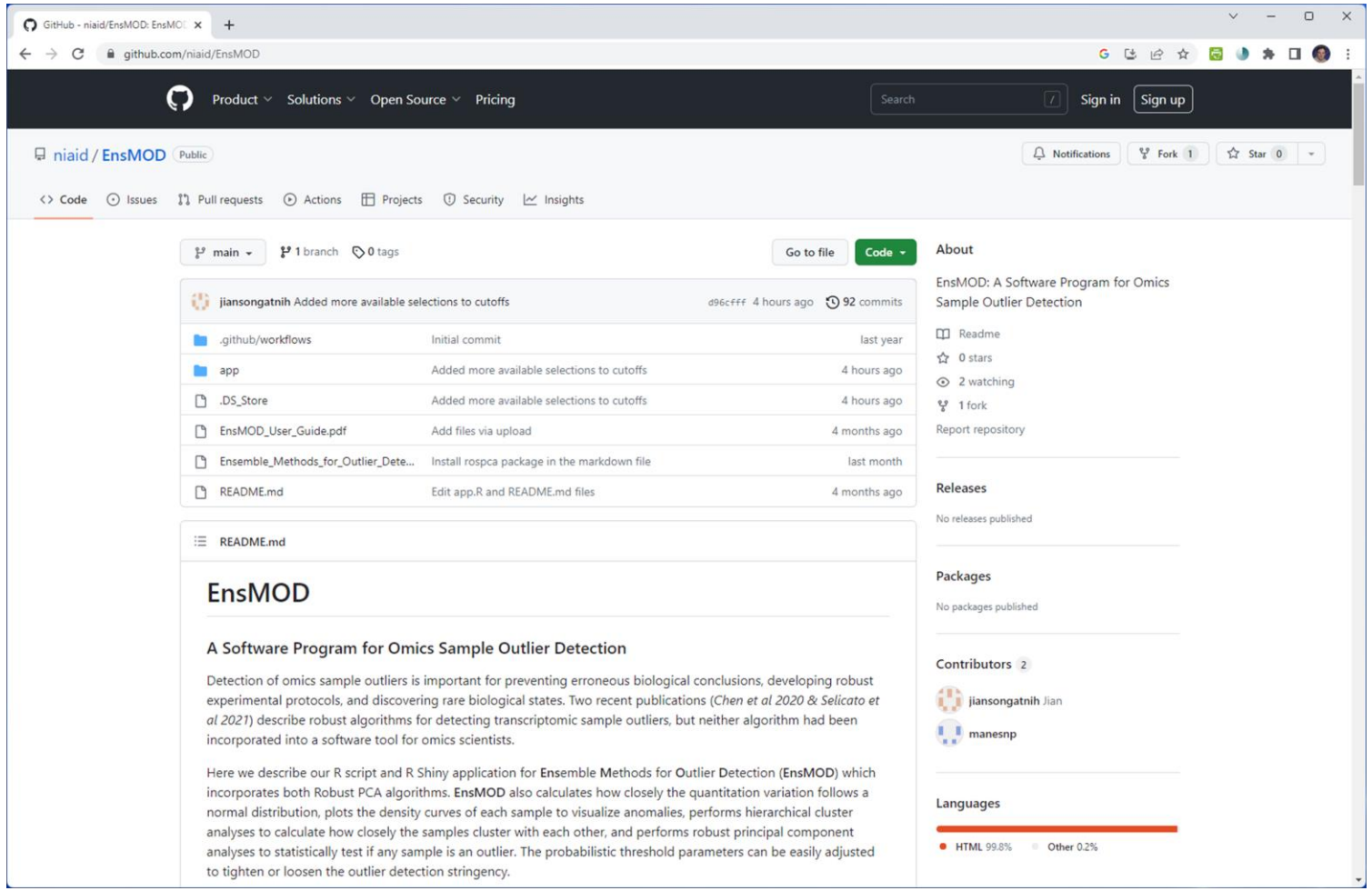
Received: 1 March 2021

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

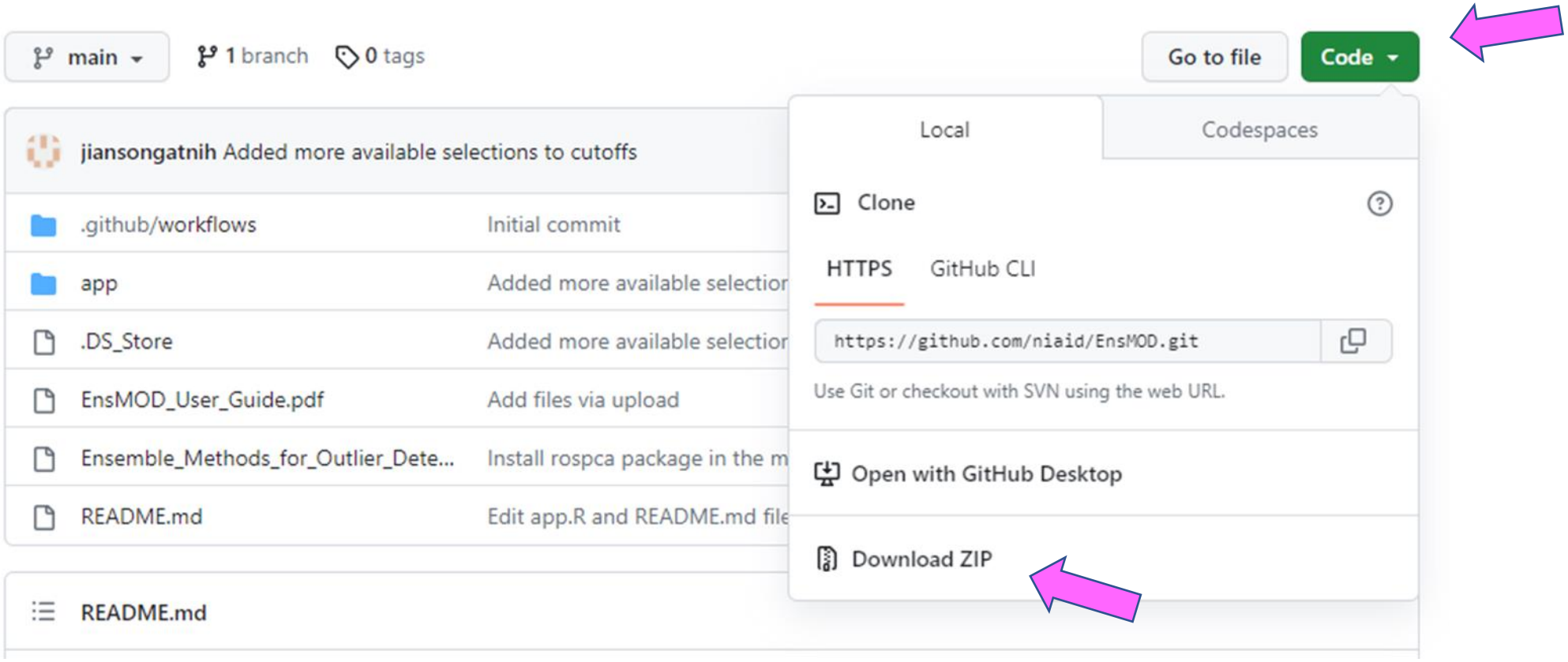


- Free and open-source
- Uses both previously published algorithms for robust transcriptomics sample outlier detection
- Can be used to analyze any omics dataset (~Gaussian variance, ≥9 samples)
- No need for a strong background in bioinformatics or biostatistics
- Step-by-step user guide
- Published

https://github.com/niaid/EnsMOD

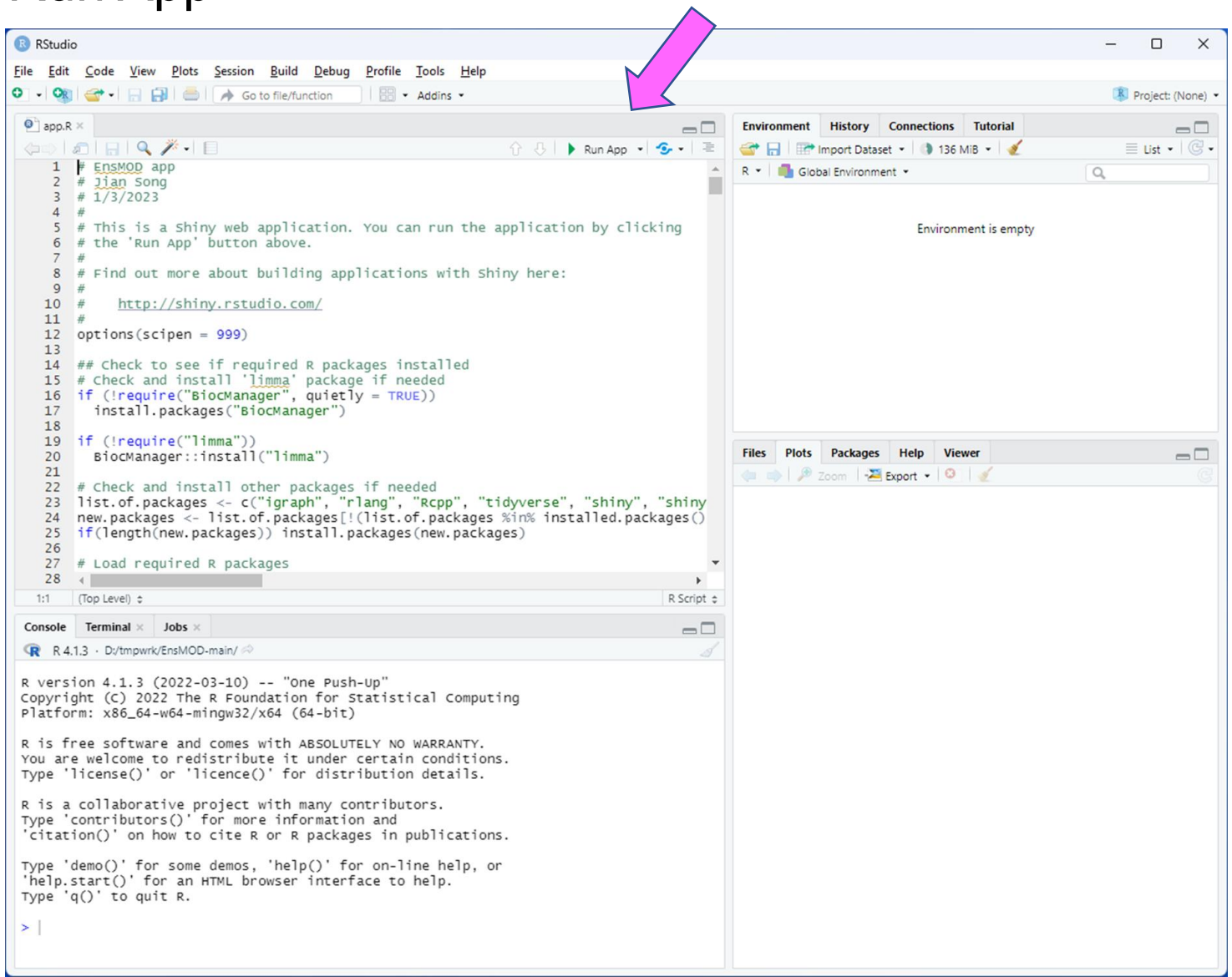
Using EnsMOD

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



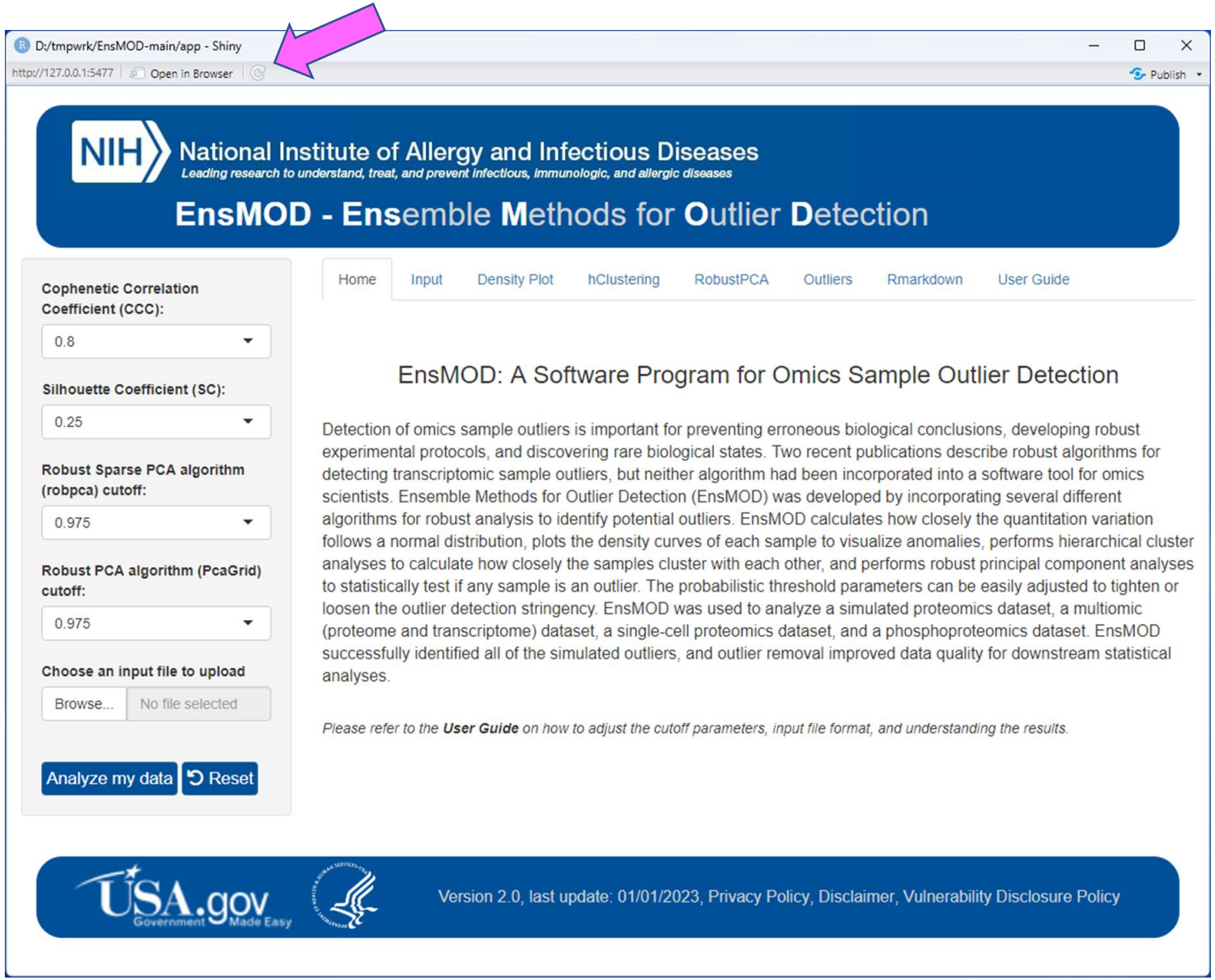
Using EnsMOD

4. Run RStudio and open /app/app.R
5. Click “Run App”



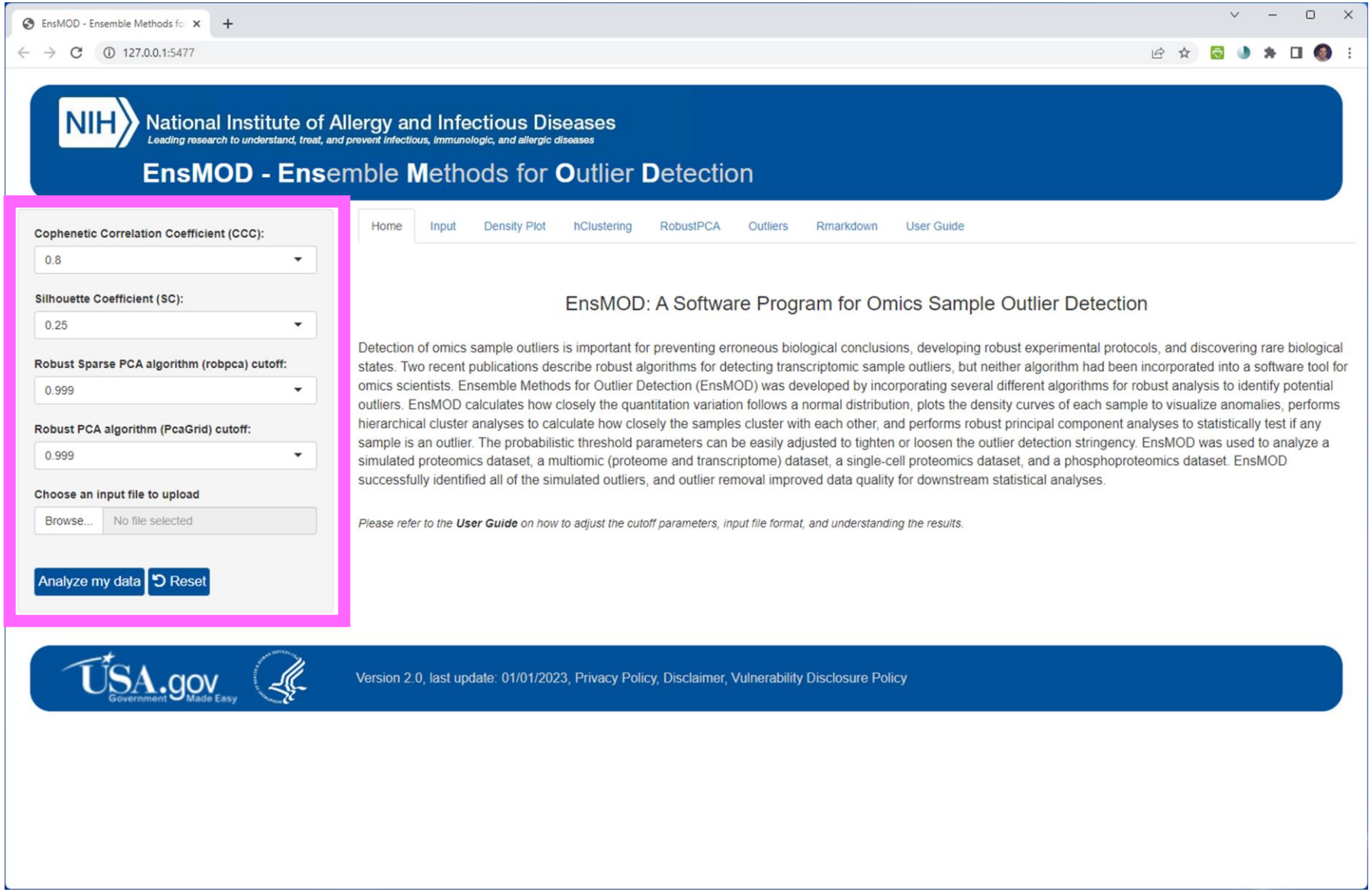
Using EnsMOD

6. Click “Open in Browser” (optional)



Using EnsMOD

7. Set parameters and select expression file



EnsMOD Input Expression File

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

Columns = samples

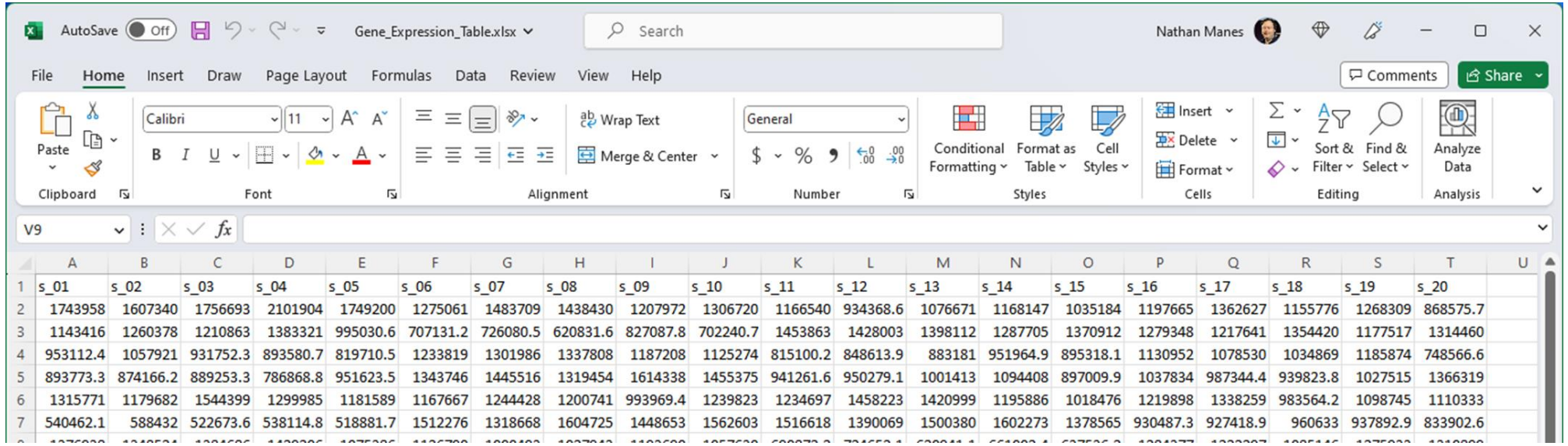
Rows = analytes

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)

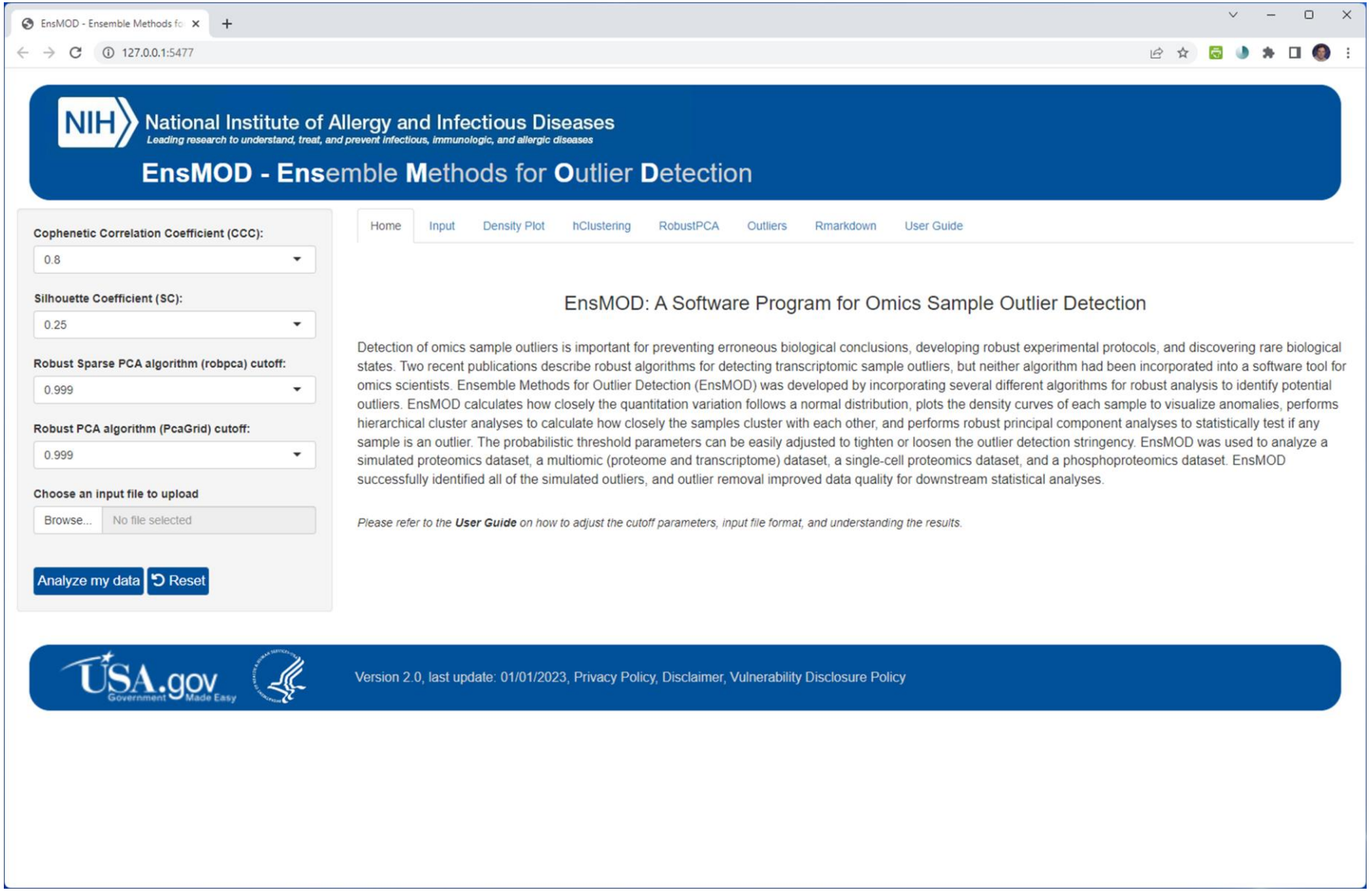
Note: Rows that contain ≥1 missing values (e.g., “NaN”) are ignored



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
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

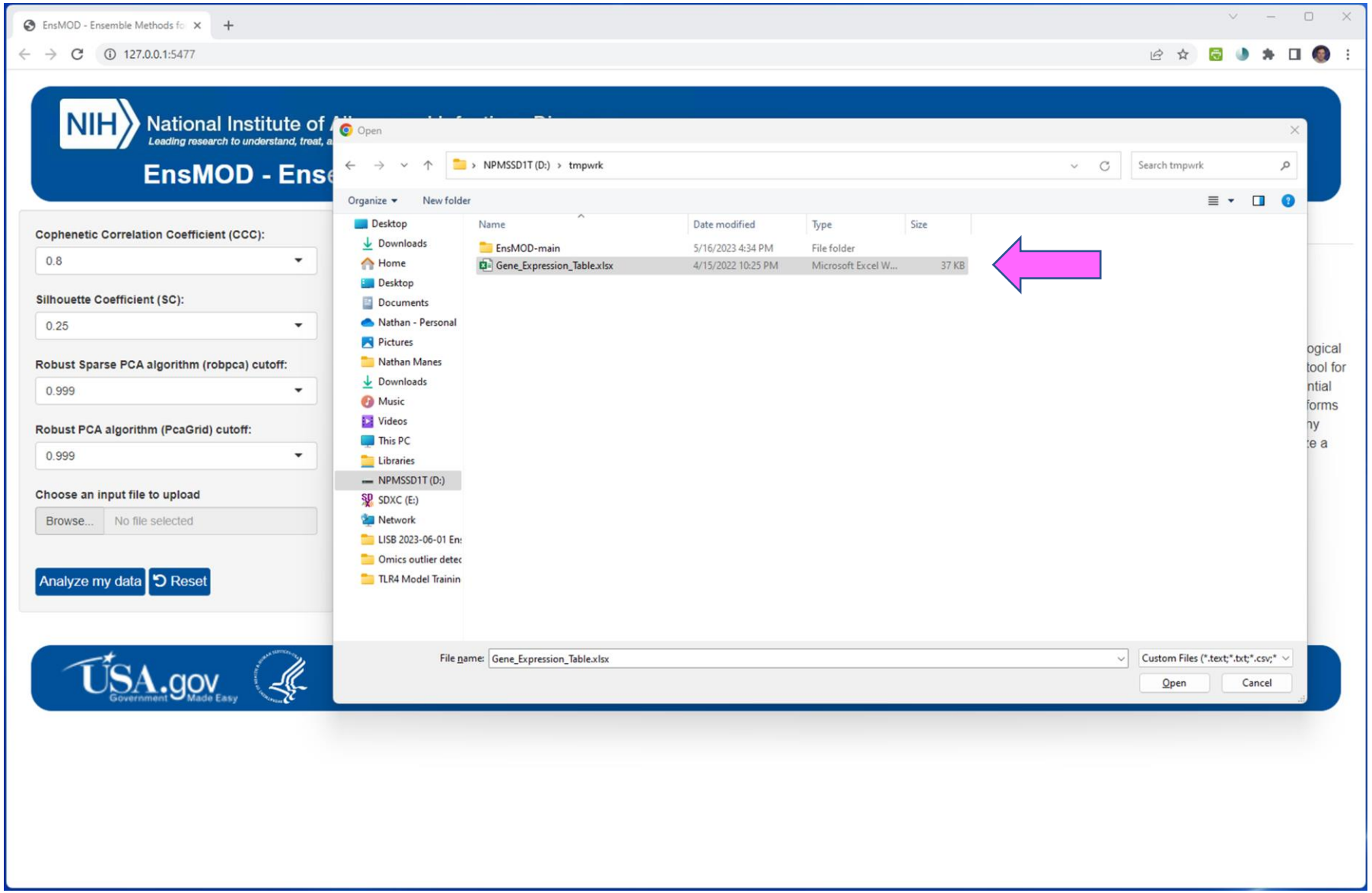
Using EnsMOD

7. Set parameters and select expression file



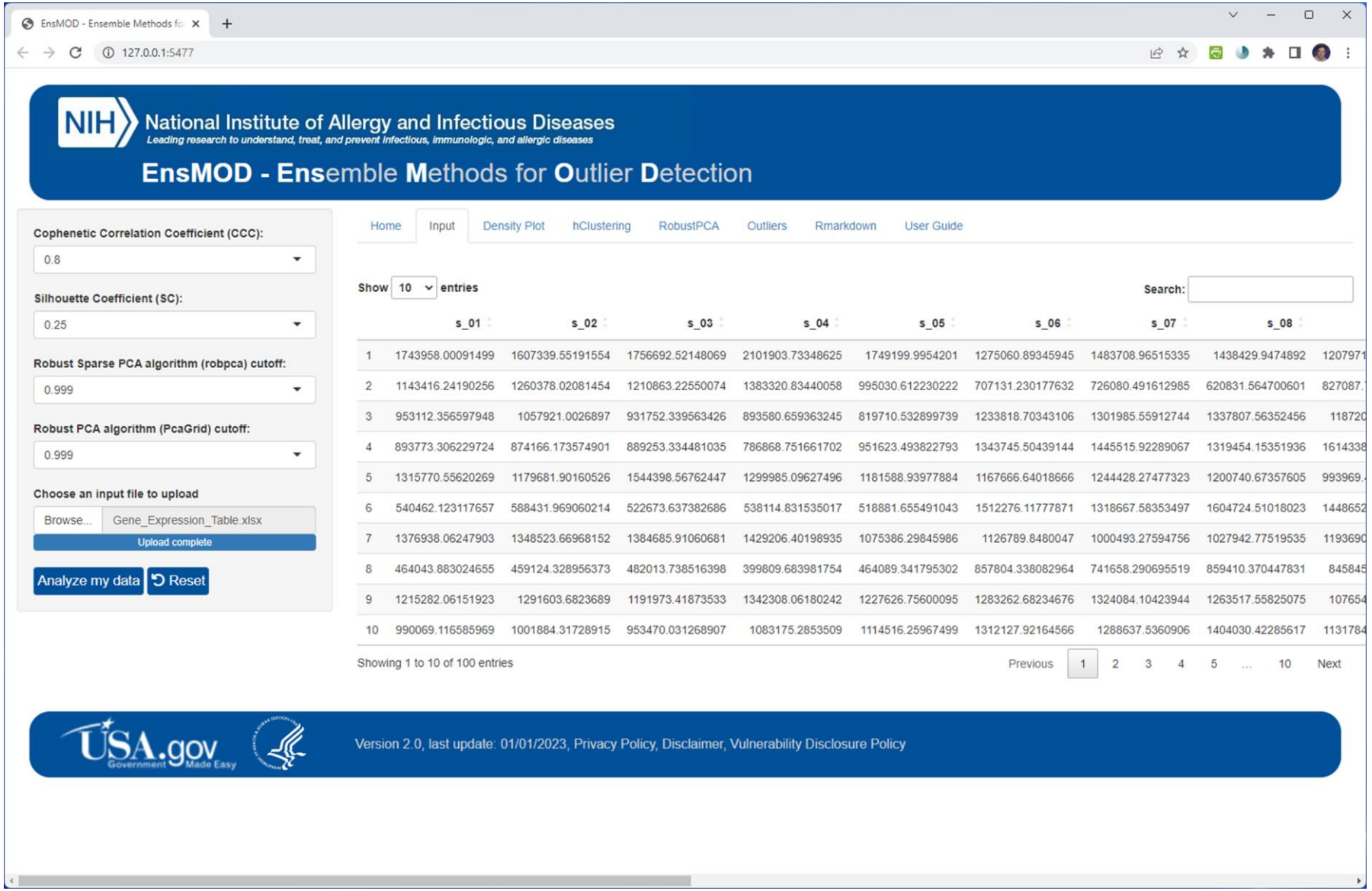
Using EnsMOD

7. Set parameters and select expression file



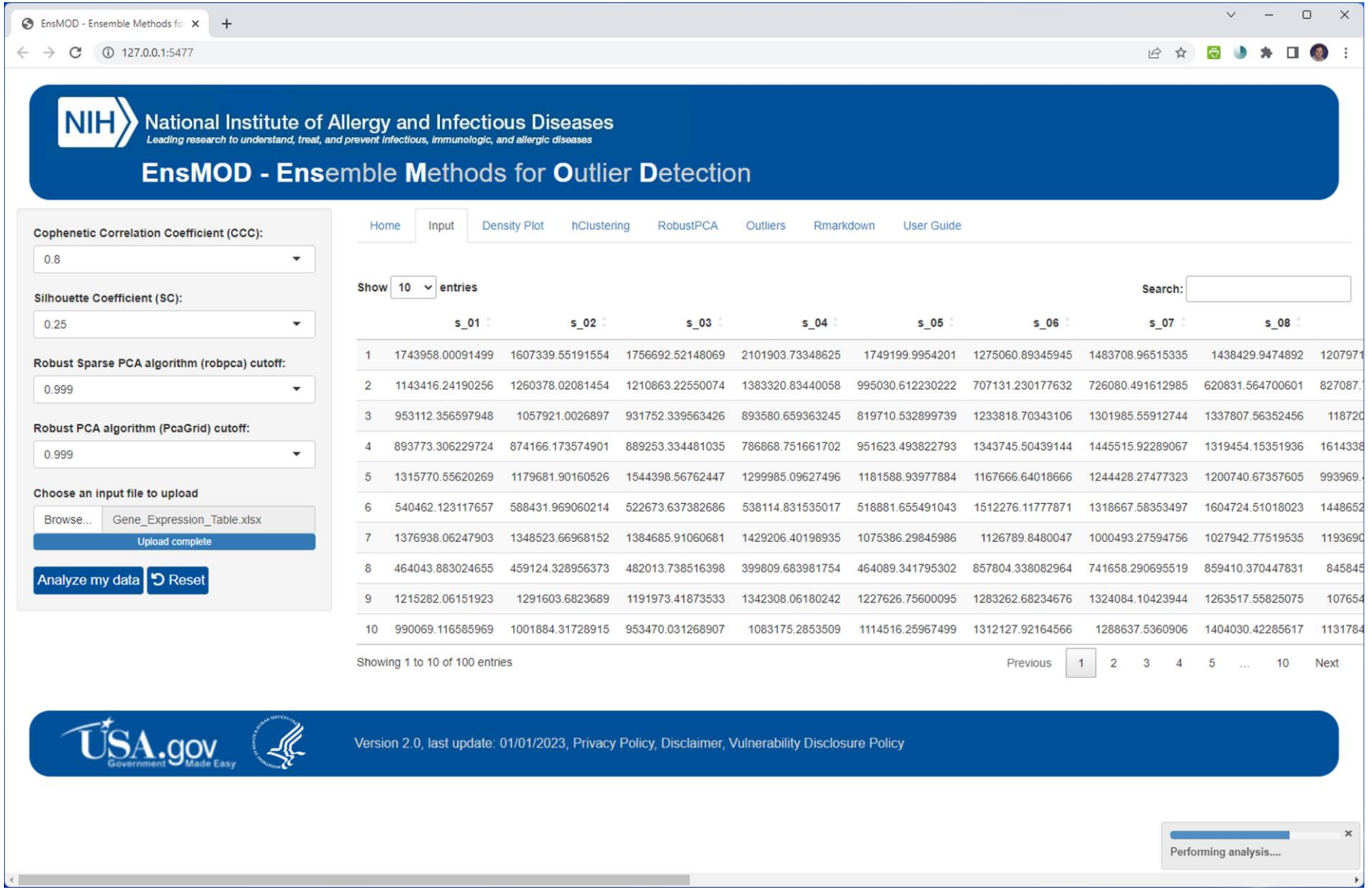
Using EnsMOD

8. Click “Analyze my data”

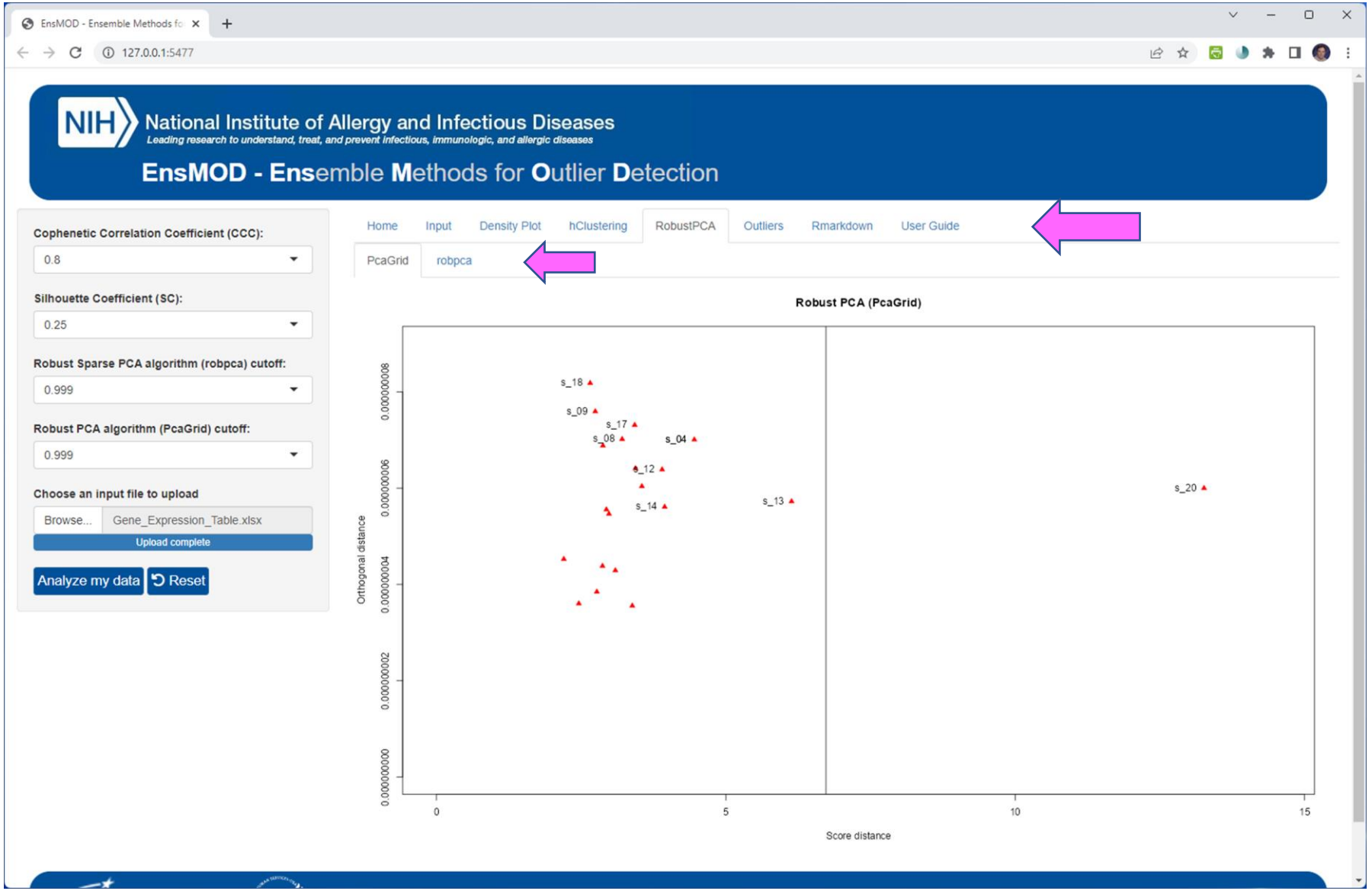


Using EnsMOD

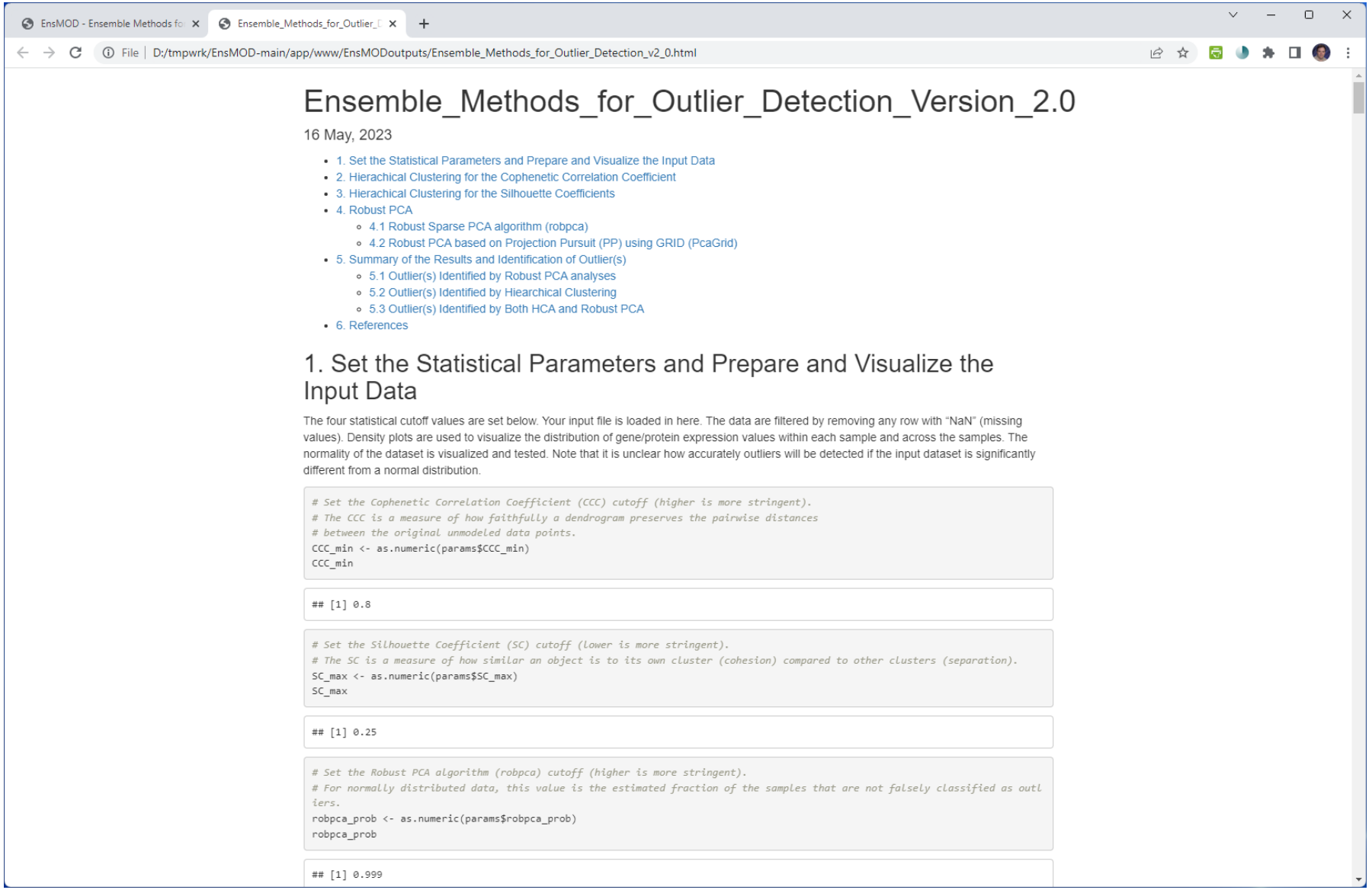
8. Click “Analyze my data”



Using EnsMOD

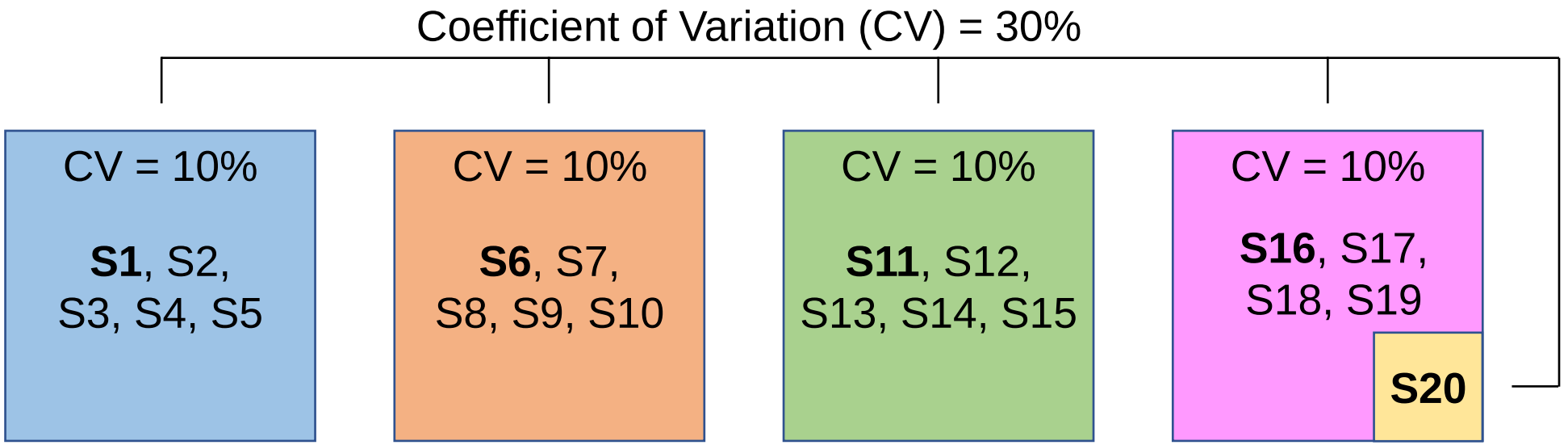


EnsMOD HTML Output File of all Results is in \app\www\EnsMODoutputs\

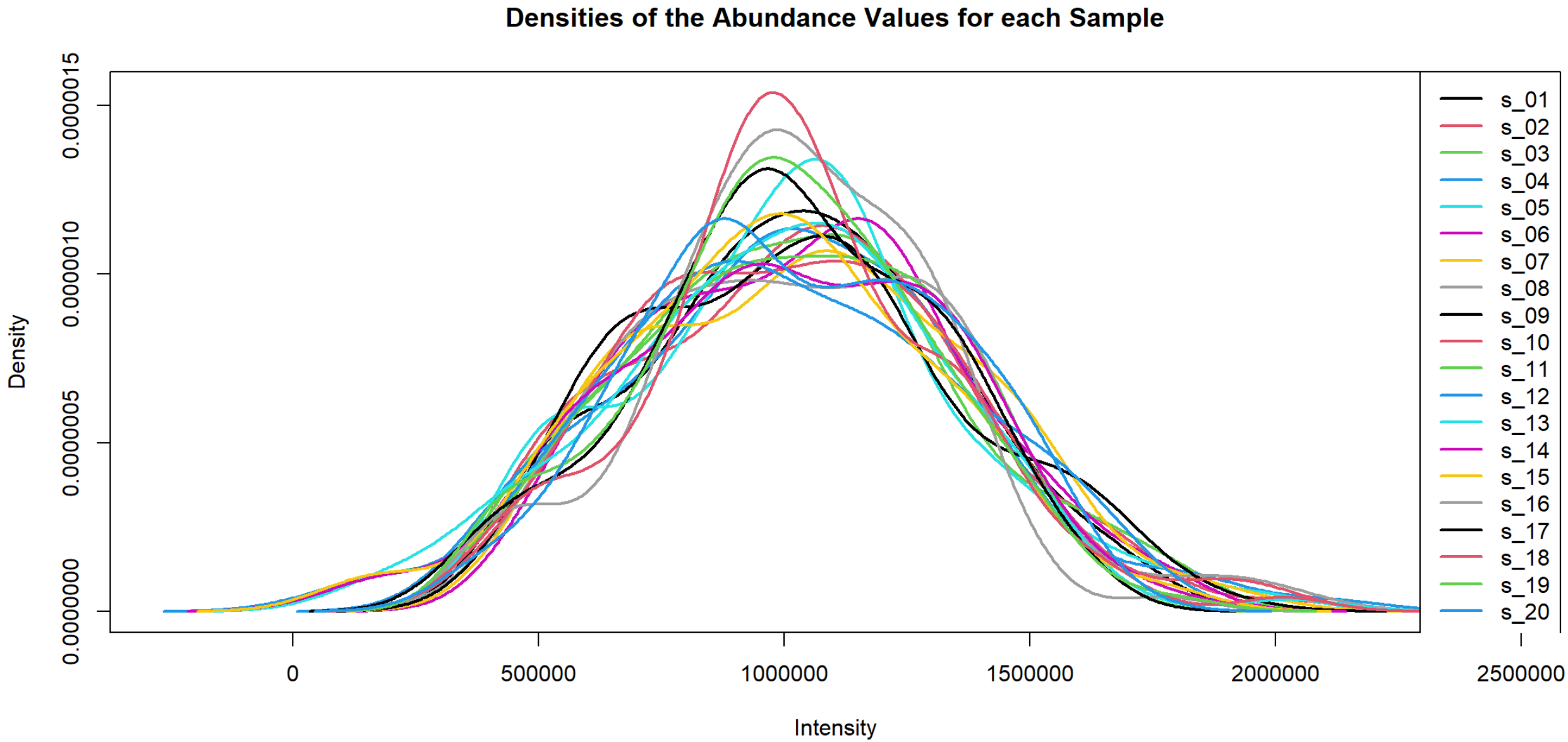


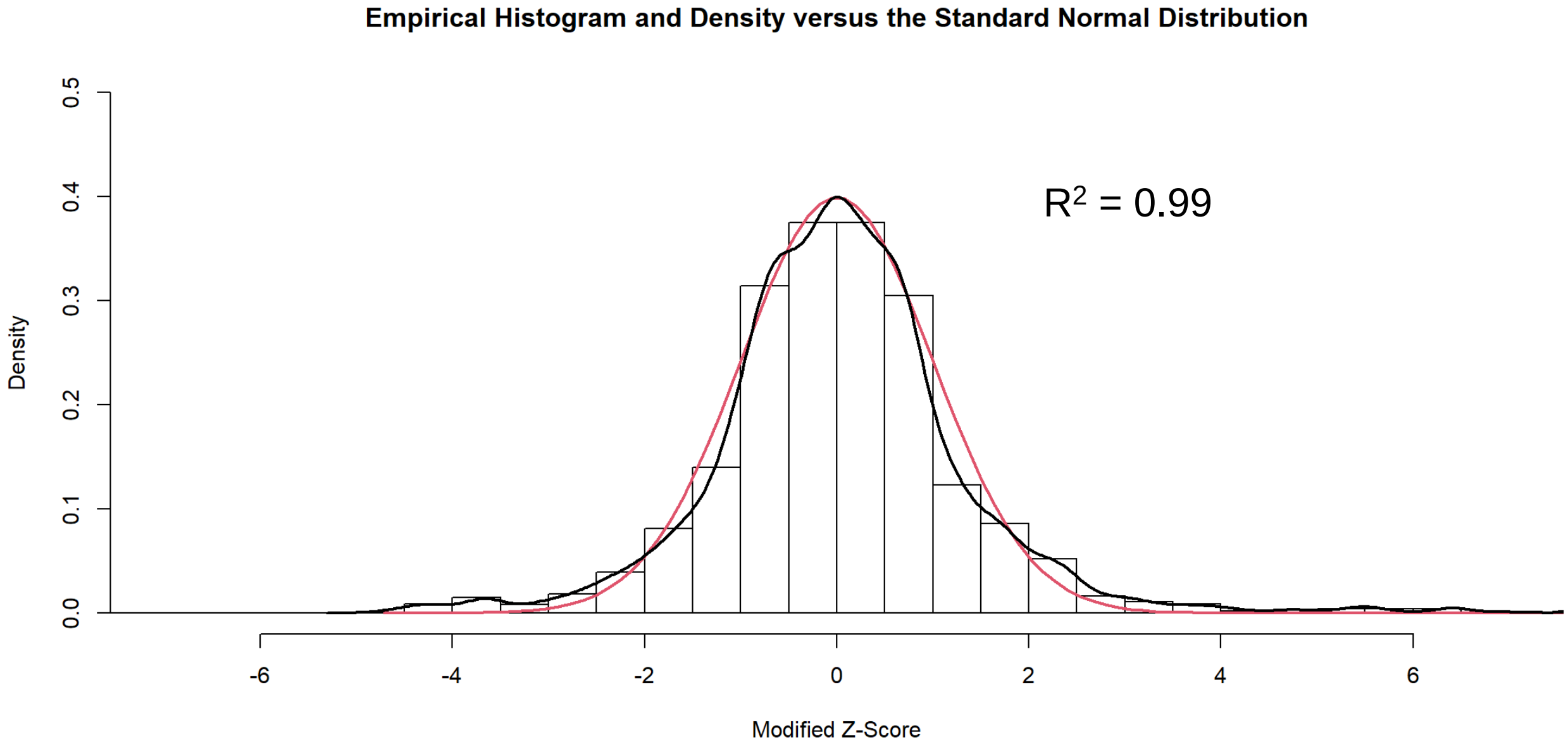
Simulated Proteomics Dataset

Simulated Proteomics Dataset



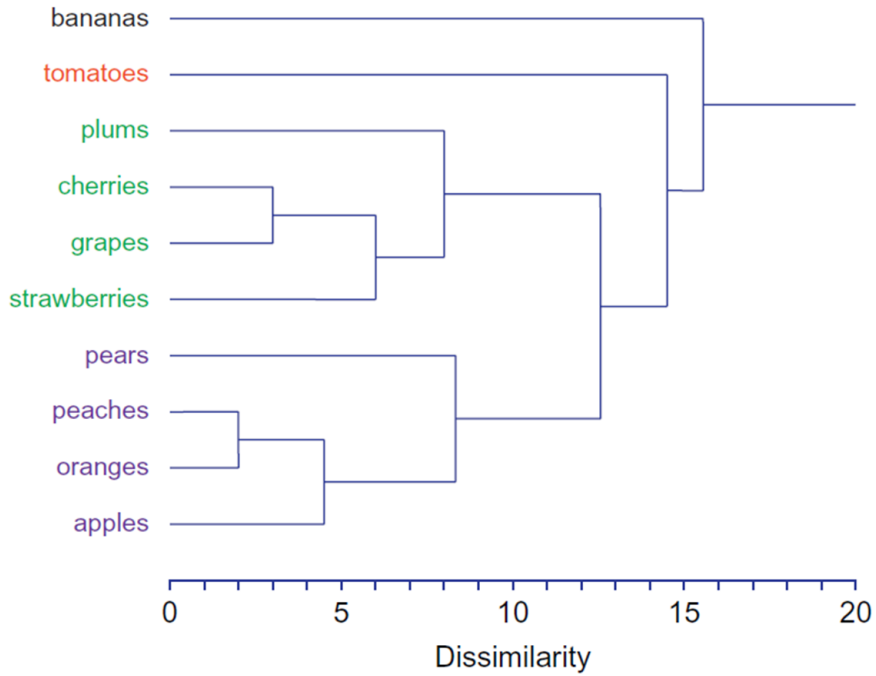
S1, S6, S11, S16, S20 = Initial Gaussian random sampling
μ = 1,000,000, CV = 30%, N=100 proteins


Densities of the Abundance Values for each Sample
Density
0.0000000
0.0000005
0.0000010
0.0000015
0
500000
1000000
1500000
2000000
2500000
Intensity
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


Empirical Histogram and Density versus the Standard Normal Distribution
$R^2 = 0.99$
Density
0.0
0.1
0.2
0.3
0.4
0.5
-6
-4
-2
0
2
4
6
Modified Z-Score

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										—



Tom Tullis, Bill Albert; Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics, 2nd Ed., 2013, Elsevier Inc.

Hierarchical Cluster Analyses and the Cophenetic Correlation Coefficient (CCC)

3 Distances: Euclidean, Manhattan, Pearson Correlation

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

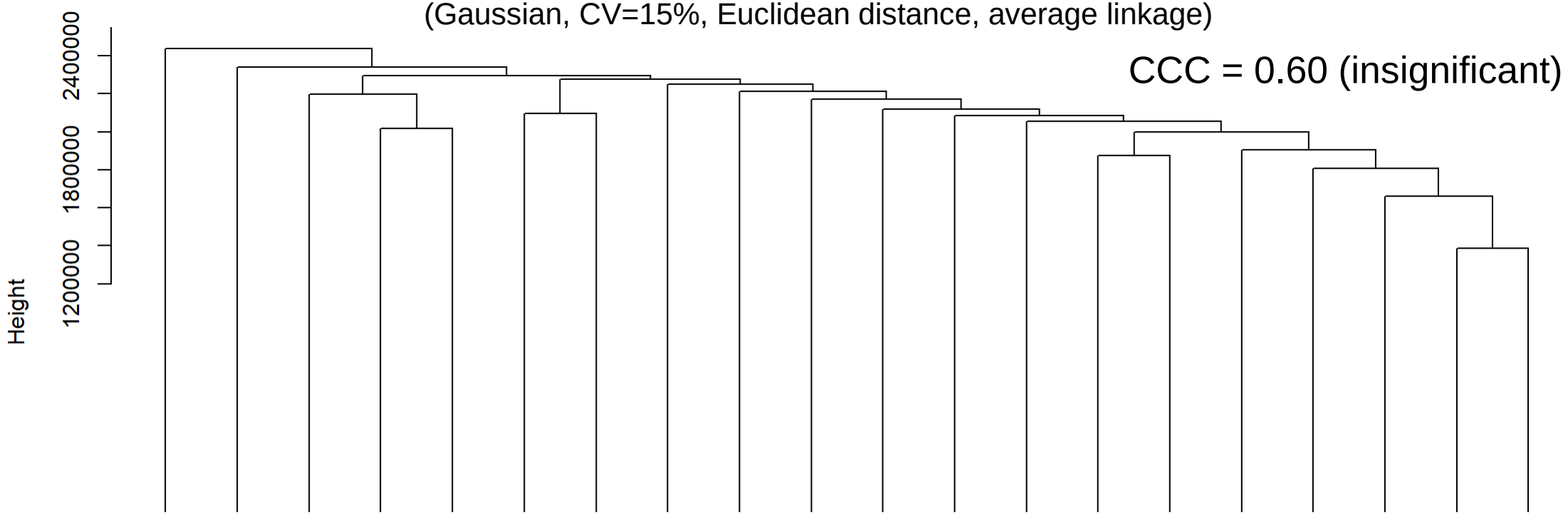
3 × 5 = 15 HCAs total

The HCA with the largest CCC is selected for downstream analyses

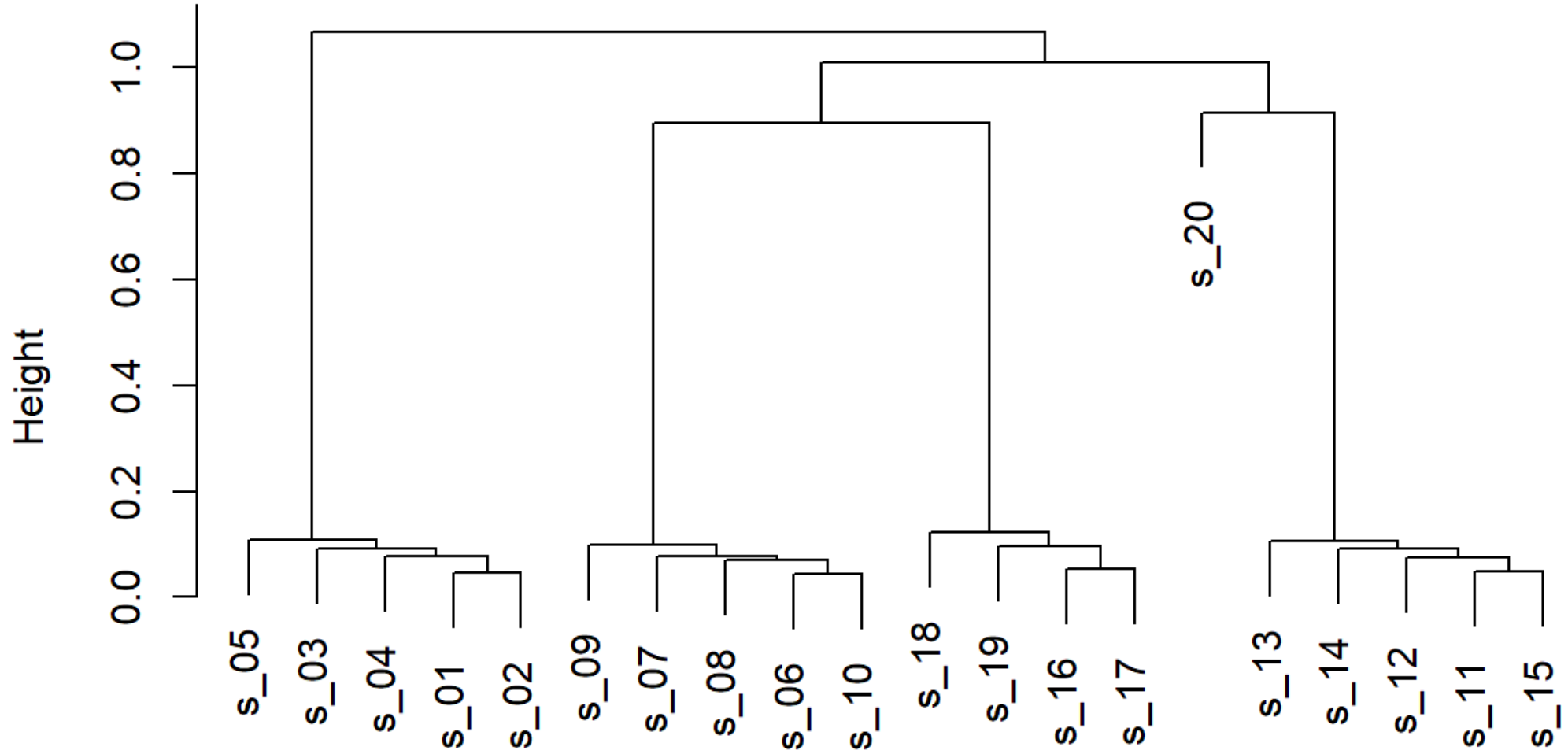
If CCC ≥ 0.8, then outliers might be identifiable [Selicato et al 2021 Mathematics 9:882]

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

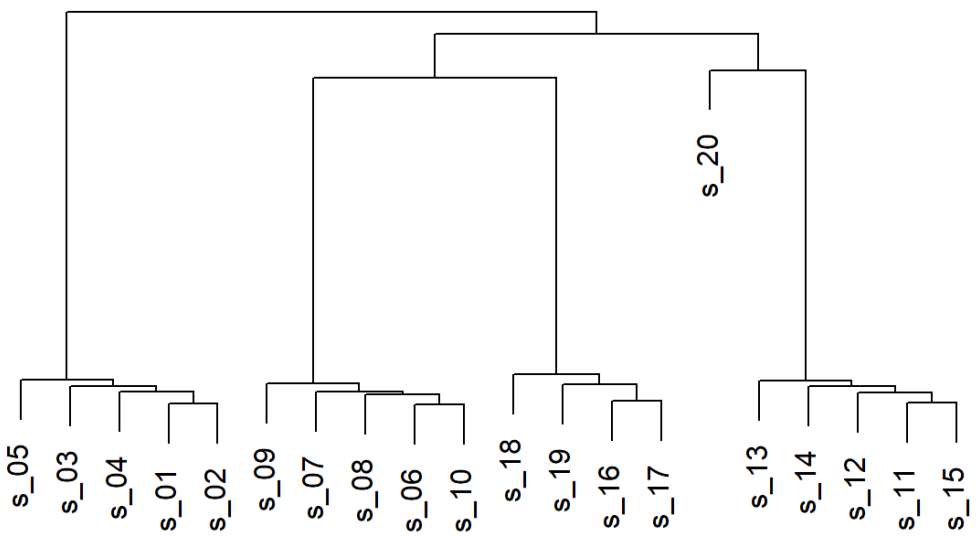
How well the sample would fit in the other HCA clusters

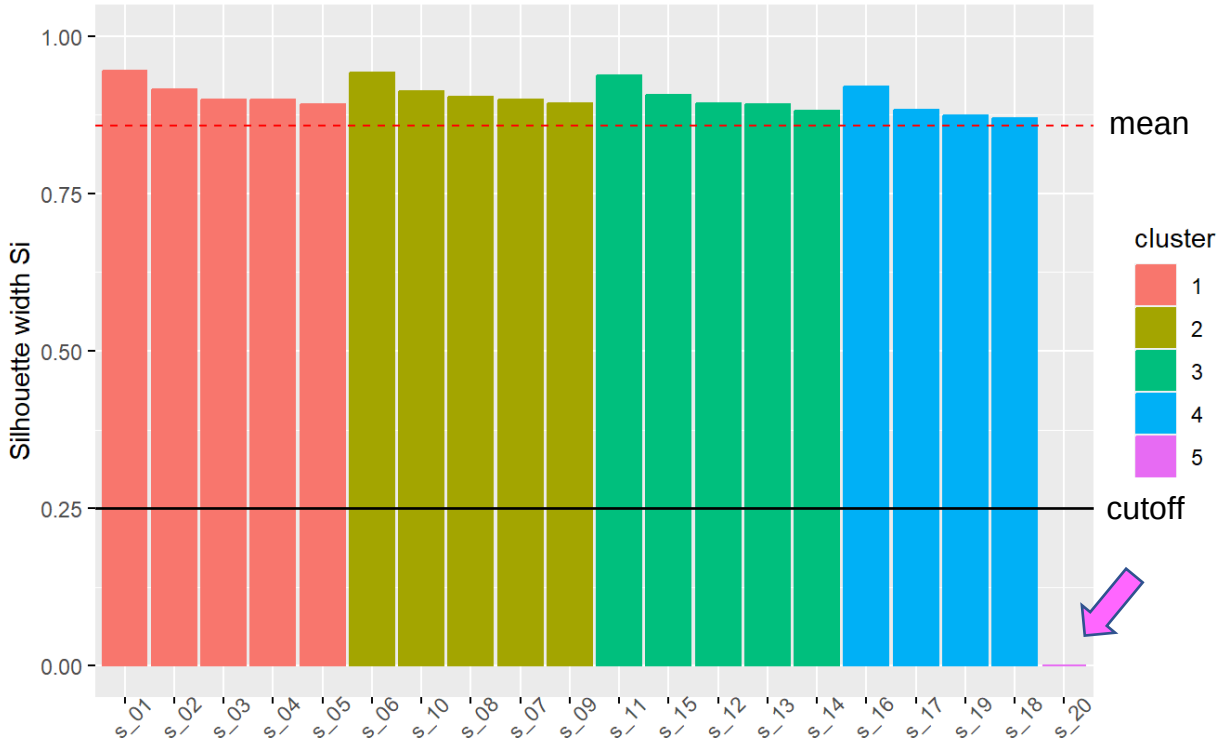
Table 1. Structures corresponding to particular Silhouette coefficient ranges.

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.

Selicato et al 2021 Mathematics 9:882

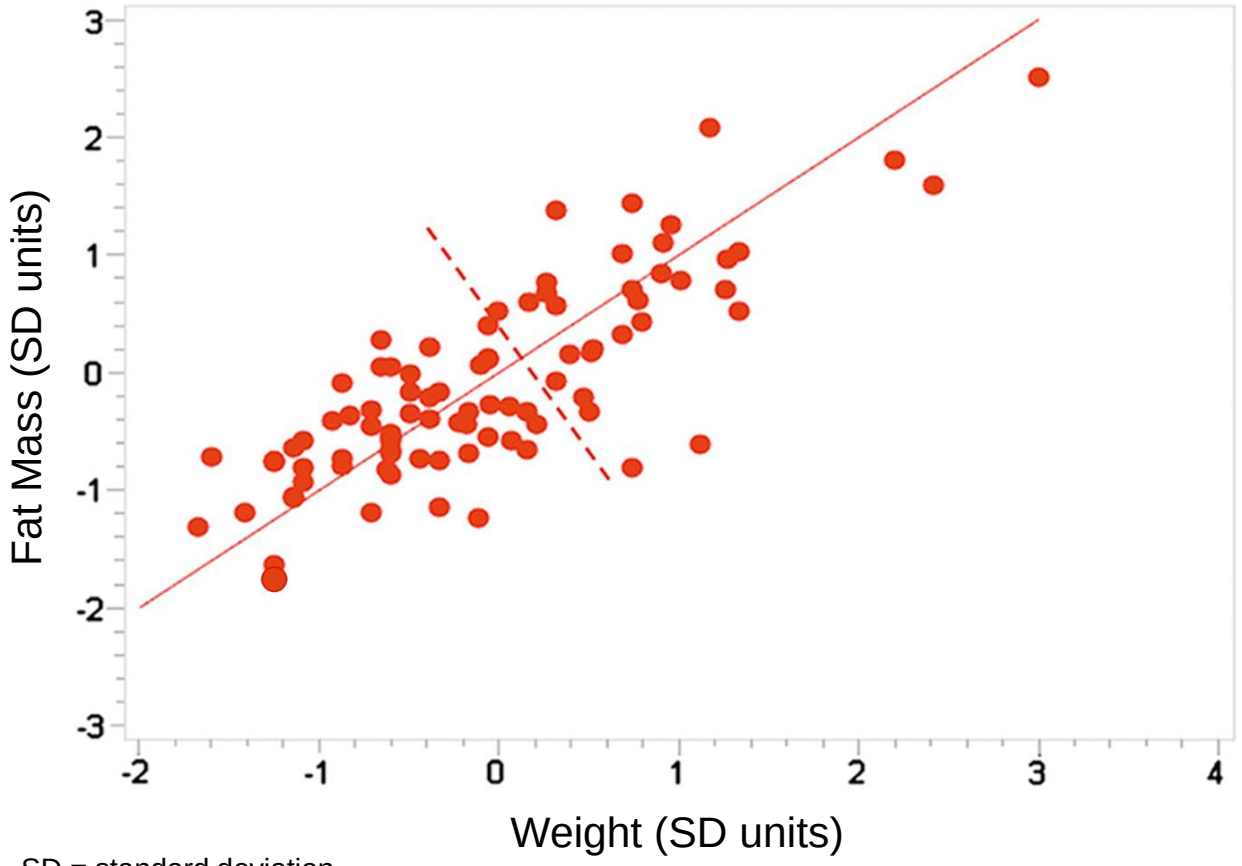
If SC < 0.25 (default parameter value), then the sample might be an outlier

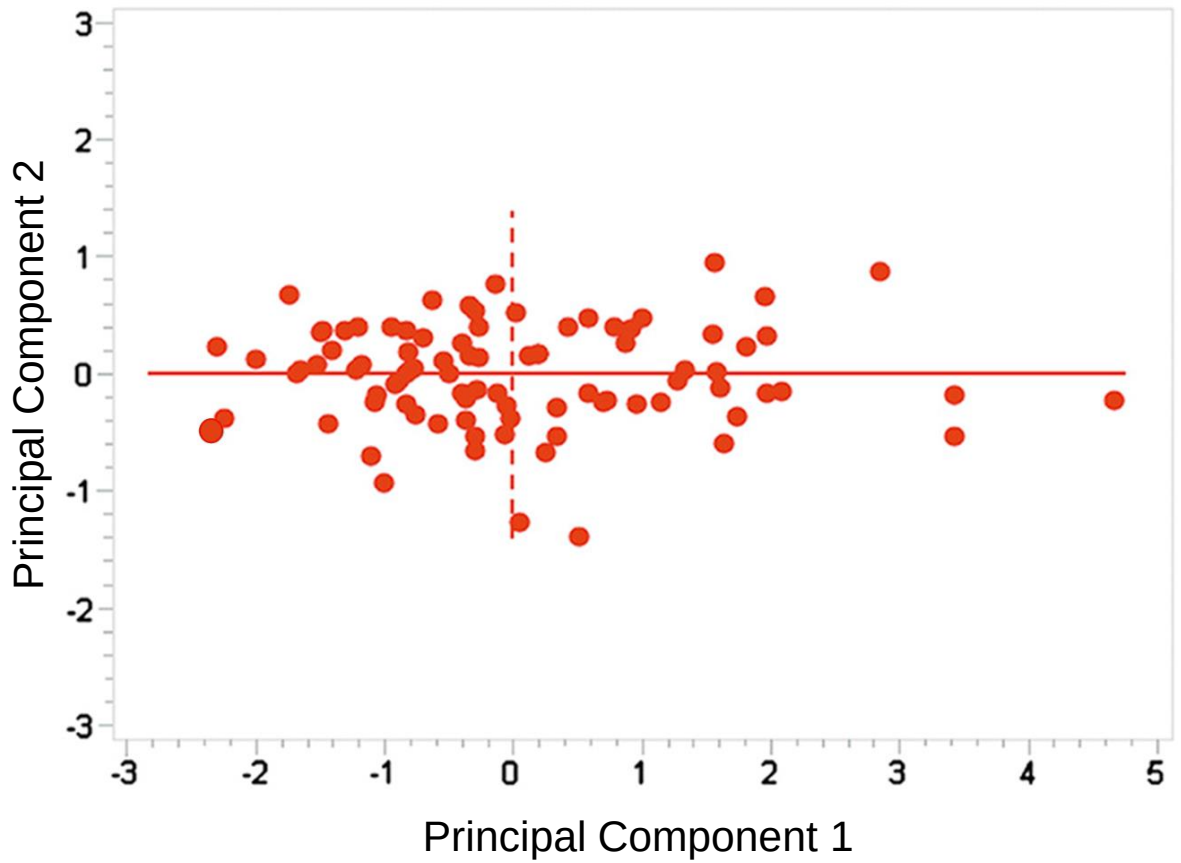




SC[s_20] = 0

Introduction to Principal Component Analysis (PCA)





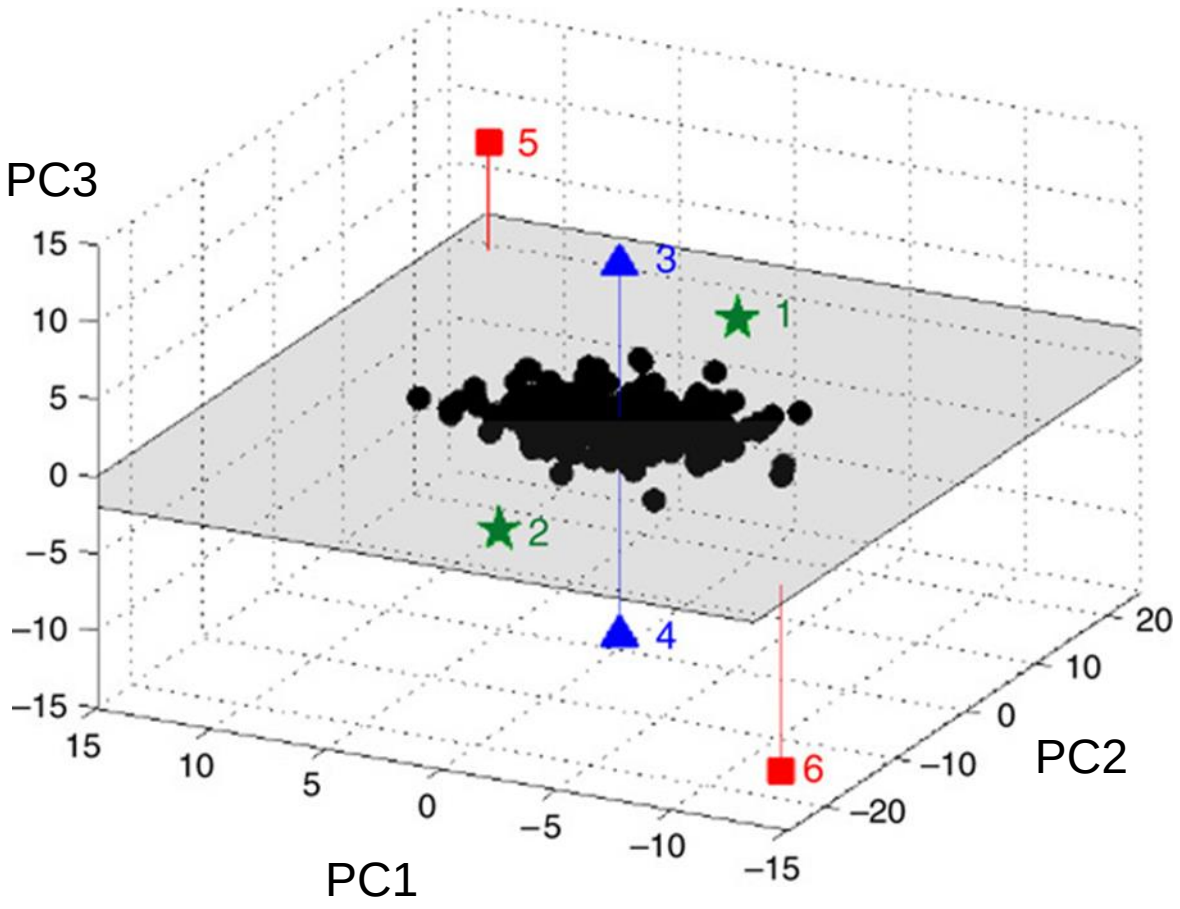
SD = standard deviation

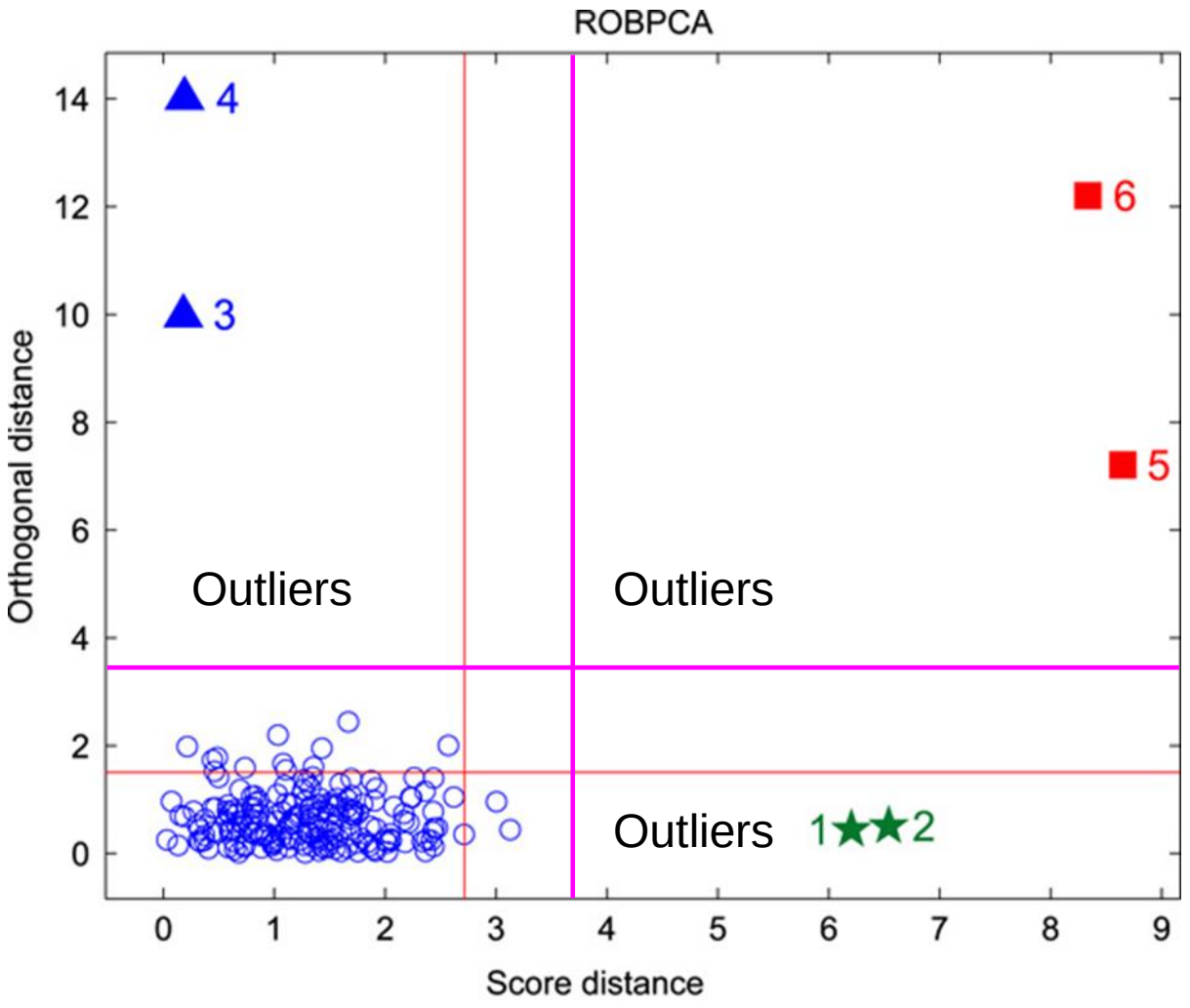
PCA: mean center and rotate for maximum variance along PC1, then PC2, ..., then PCn

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

Sainani 2014 PM&R 6:275-8

Example of a PCA Distance-Distance Plot



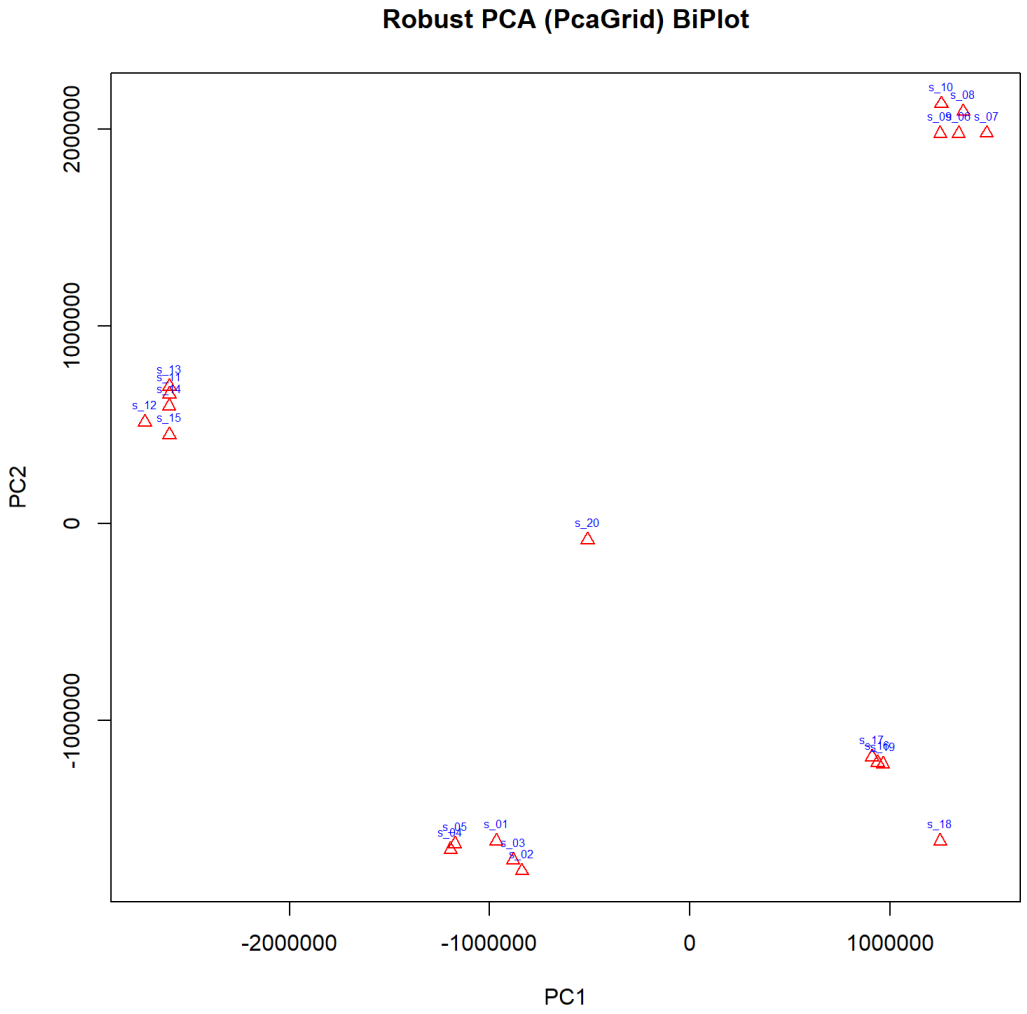


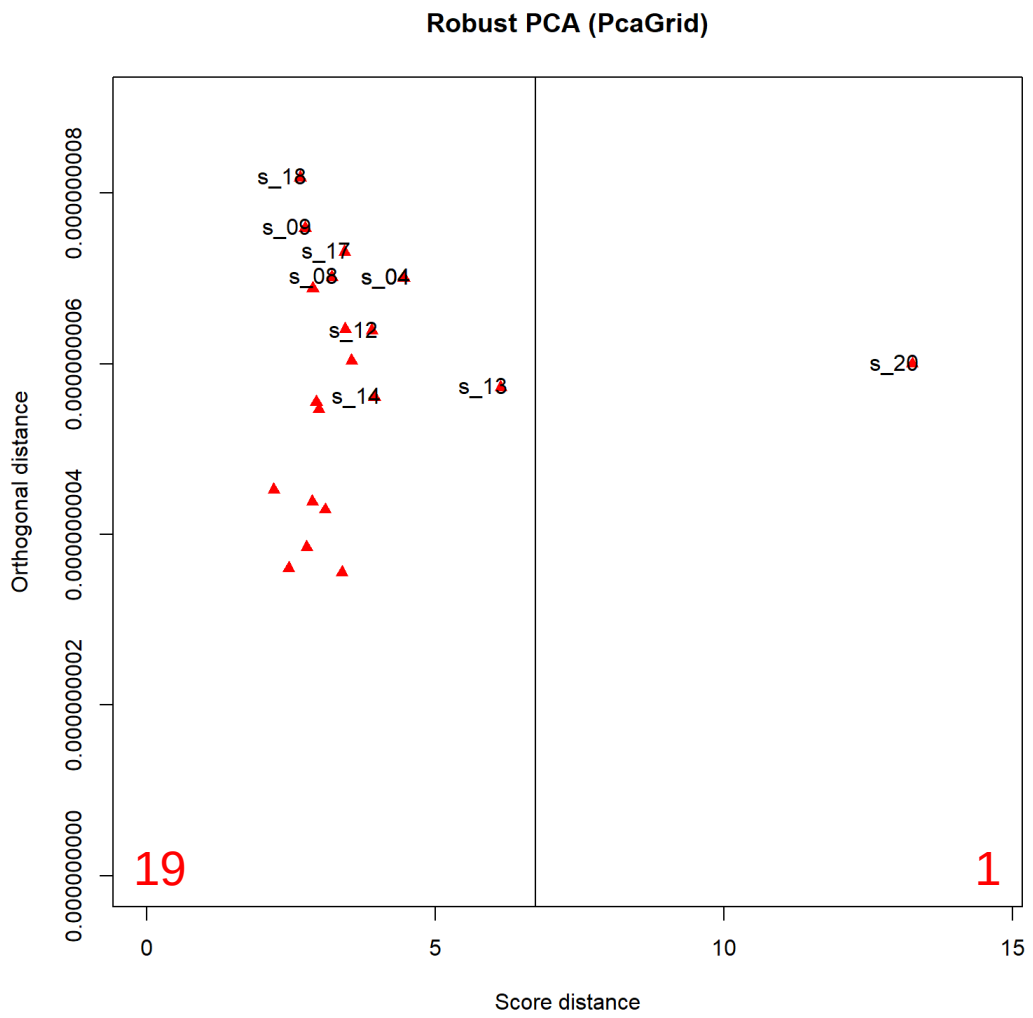
rPCA parameter: 97.5%. Est. Confidence = $(97.5\%)^2 = 95.1\%$

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

Hubert et al 2009 Comp Stat Data Anal 53:2264-74

Two Robust Principal Component Analyses are Performed

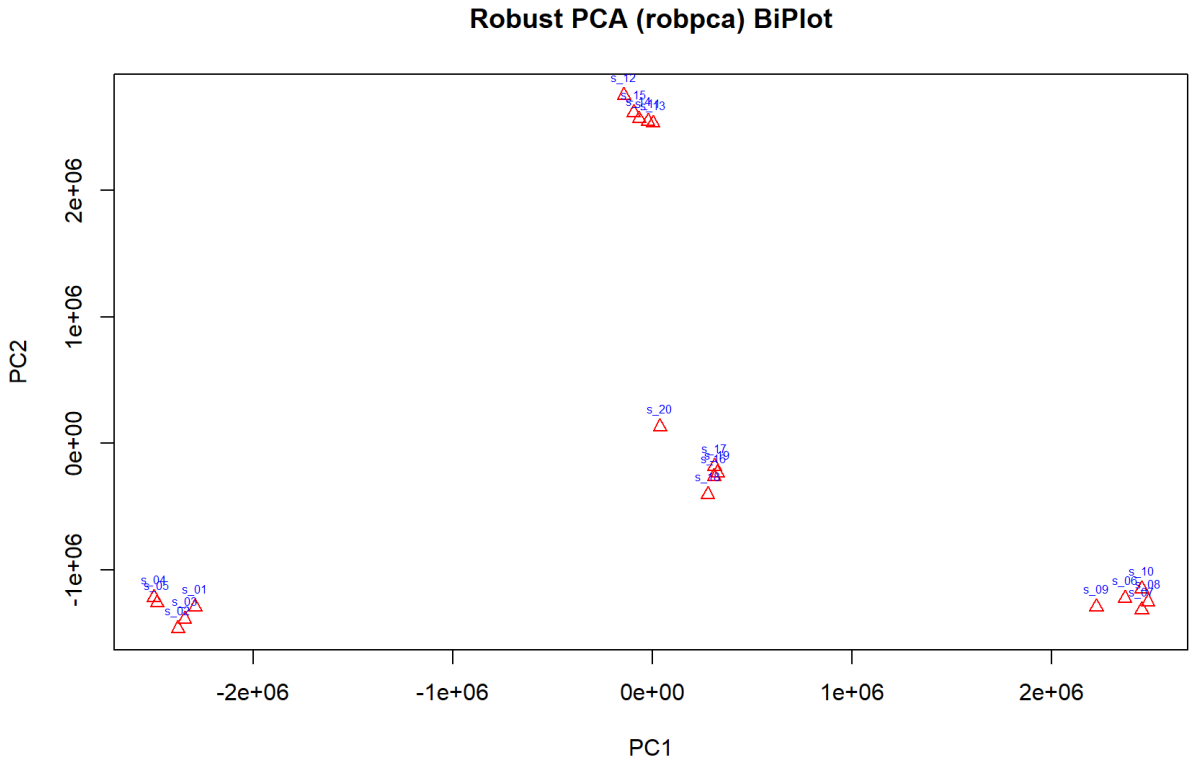


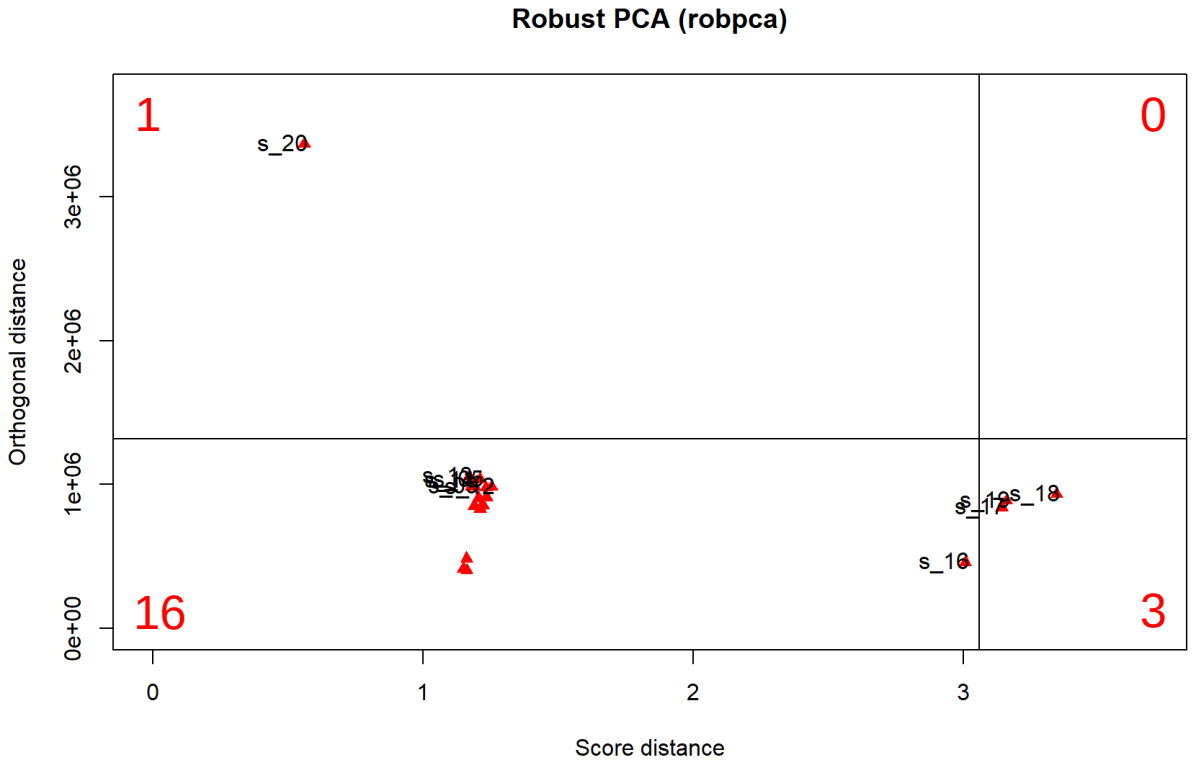


rPCA parameter was set to 99.9%

Estimated Confidence = $(99.9\%)^2 = 99.8\%$

Two Robust Principal Component Analyses are Performed





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"
```

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

BMC Bioinformatics

RESEARCH ARTICLE Open Access

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

Xiaoying Chen[1], Bo Zhang[2], Ting Wang[3,4], Azad Bonni[1] and Guoyan Zhao[1*]

* Correspondence: gzhao@wustl.edu
[1]Department of Neuroscience, Washington University School of Medicine, St. Louis, MO, USA

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%

mathematics MDPI

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]

HCA CCC ≥ 0.8

HCA SC < 0.25

robpca
p = 97.5%
Recommend: p = 99.9%

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

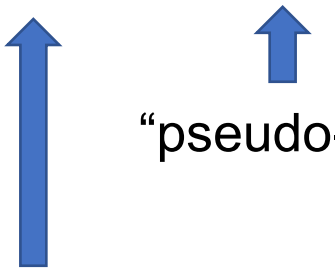
Spleens homogenized, trypsin, phosphopeptides enriched, LC-MS(/MS), Log10 label-free quantitation

Manes et al 2011 Mol Cell Proteomics 10:10.1074/mcp.M110.000927

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

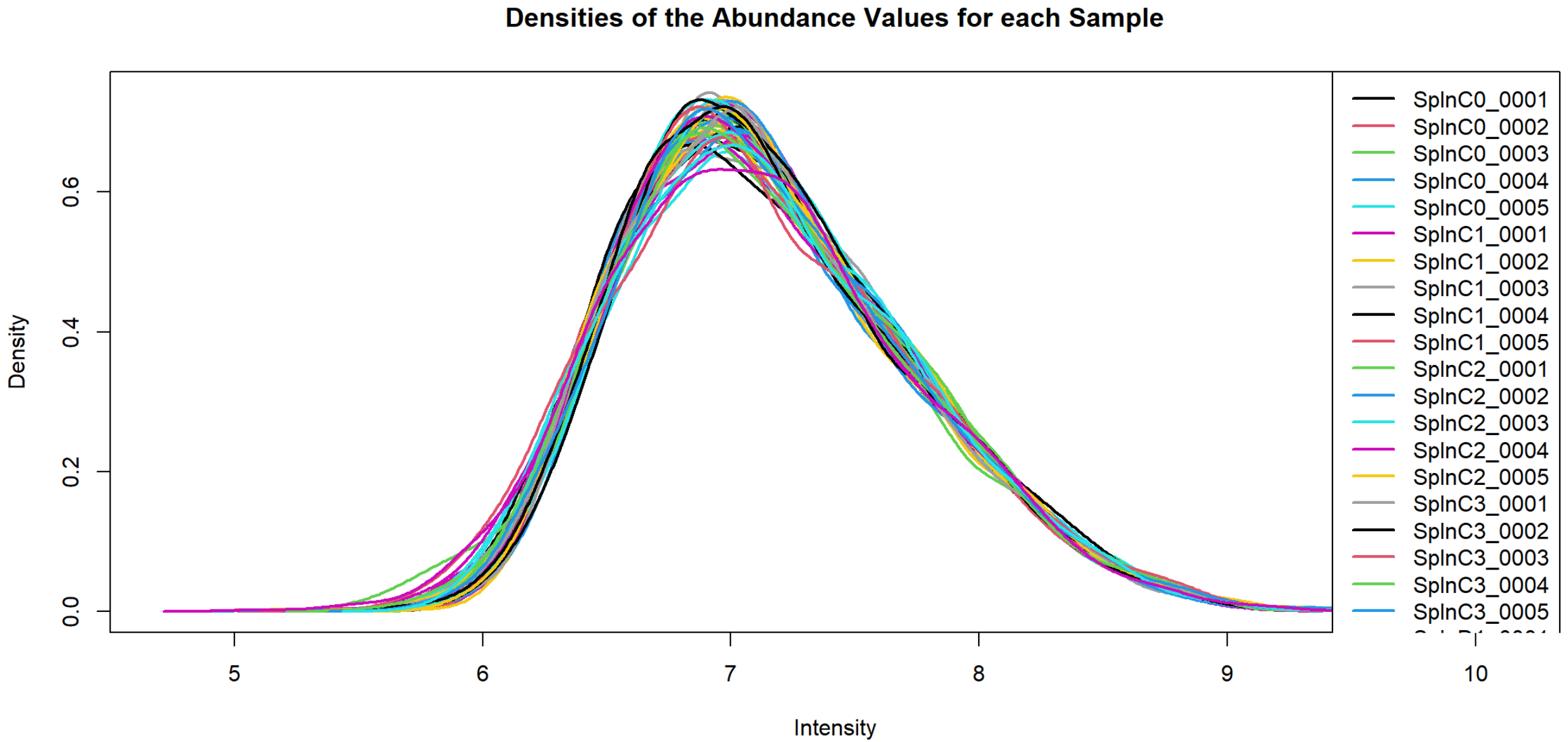
```
# The samples that satisfied all four criteria (CCC, SC, robpca, PcaGrid) for an outlier:
if (CCC_df_ranked_top$CCC >= CCC_min) {
  intersect(intersect(hcOutliers, rosOutliers), pcOutliers)
}
```

```
## [1] "SplnC0_0003" "SplnS3_0001"
```



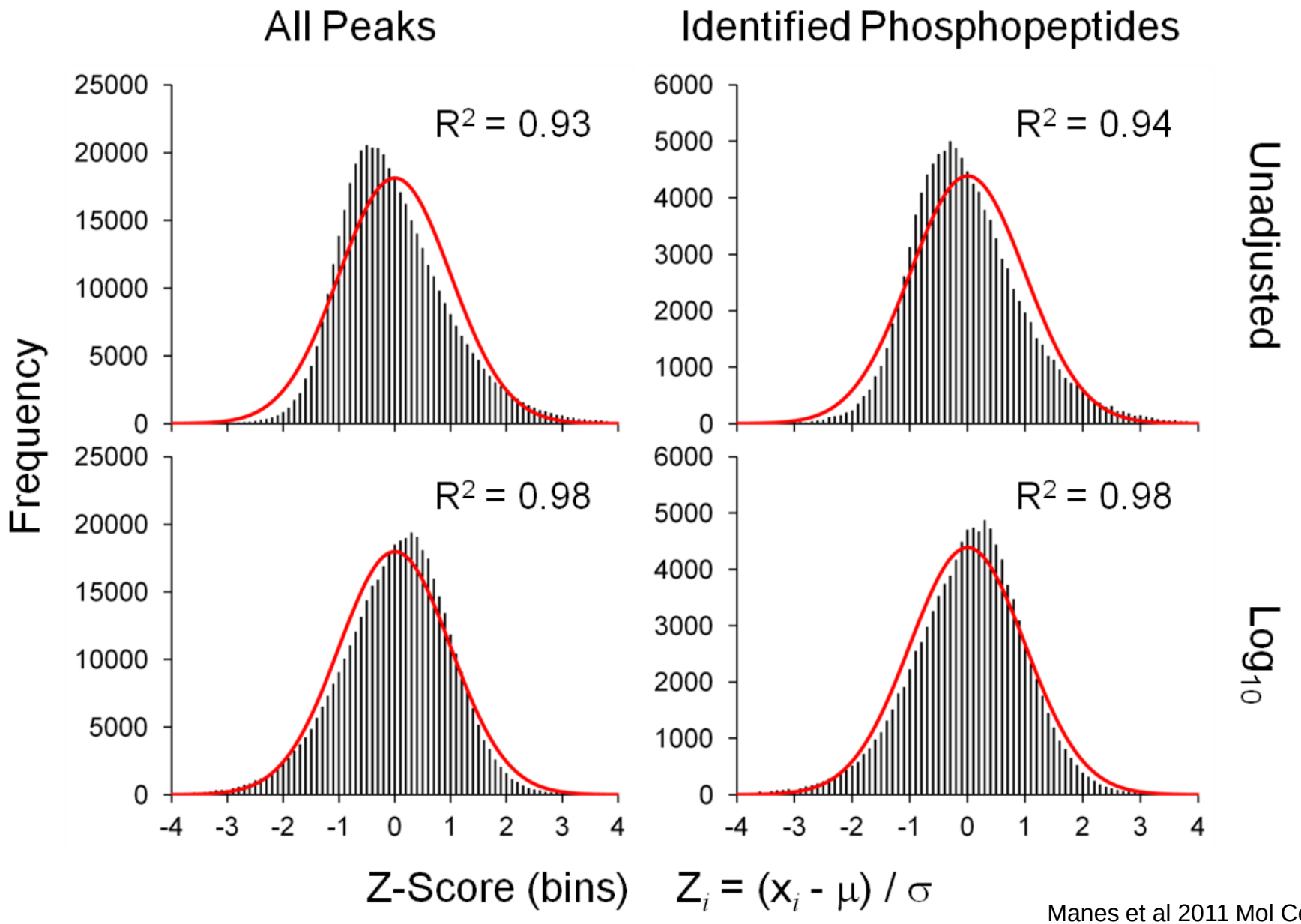
“pseudo-outlier”

One of the 20 uninfected samples

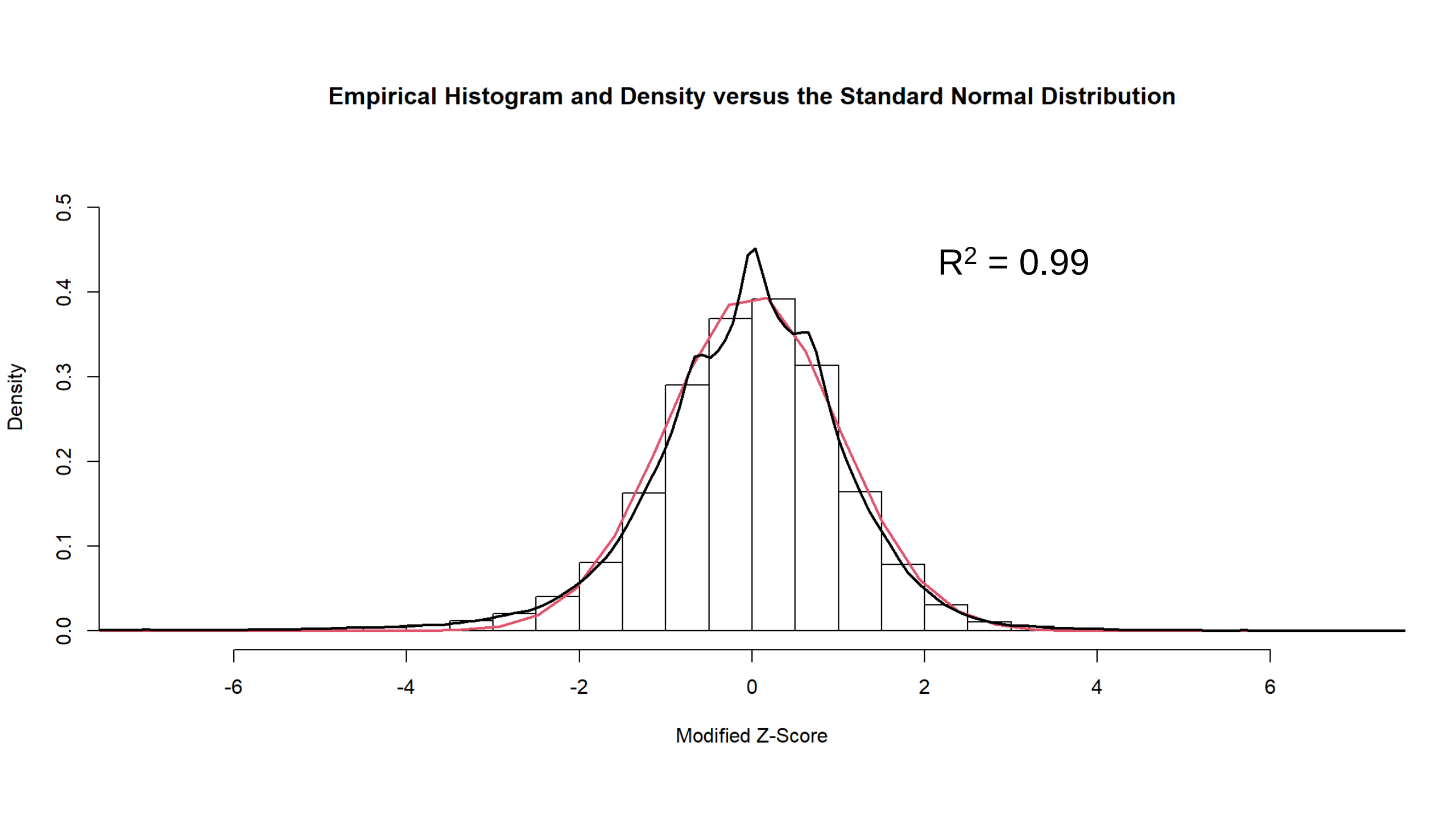

Densities of the Abundance Values for each Sample
Density
0.0
0.2
0.4
0.6
5
6
7
8
9
10
Intensity
SplnC0_0001
SplnC0_0002
SplnC0_0003
SplnC0_0004
SplnC0_0005
SplnC1_0001
SplnC1_0002
SplnC1_0003
SplnC1_0004
SplnC1_0005
SplnC2_0001
SplnC2_0002
SplnC2_0003
SplnC2_0004
SplnC2_0005
SplnC3_0001
SplnC3_0002
SplnC3_0003
SplnC3_0004
SplnC3_0005

Example of Data Transformation to Increase Normality

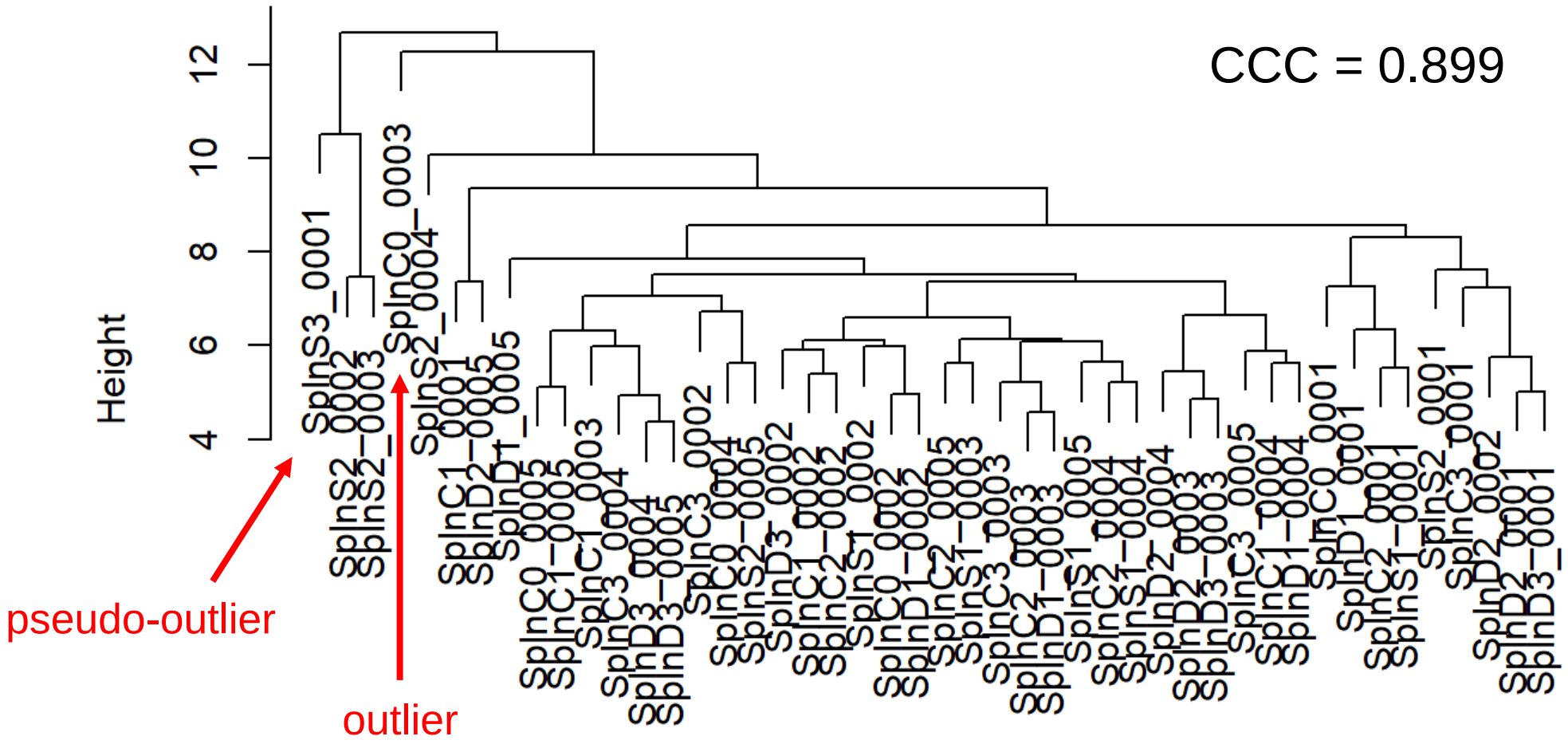
LC-MS Intensity Data are Log-Normal



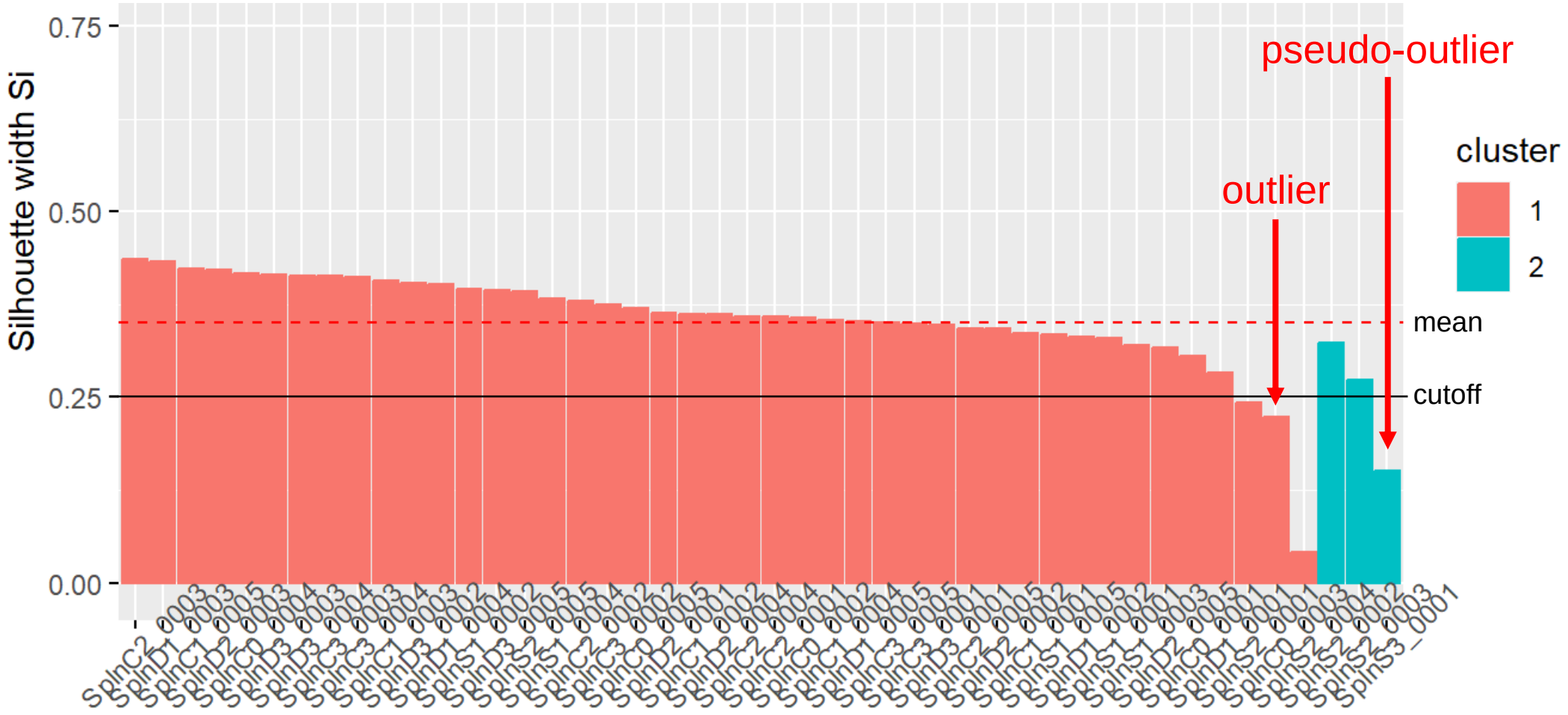
Manes et al 2011 Mol Cell Proteomics 10:M110.000927

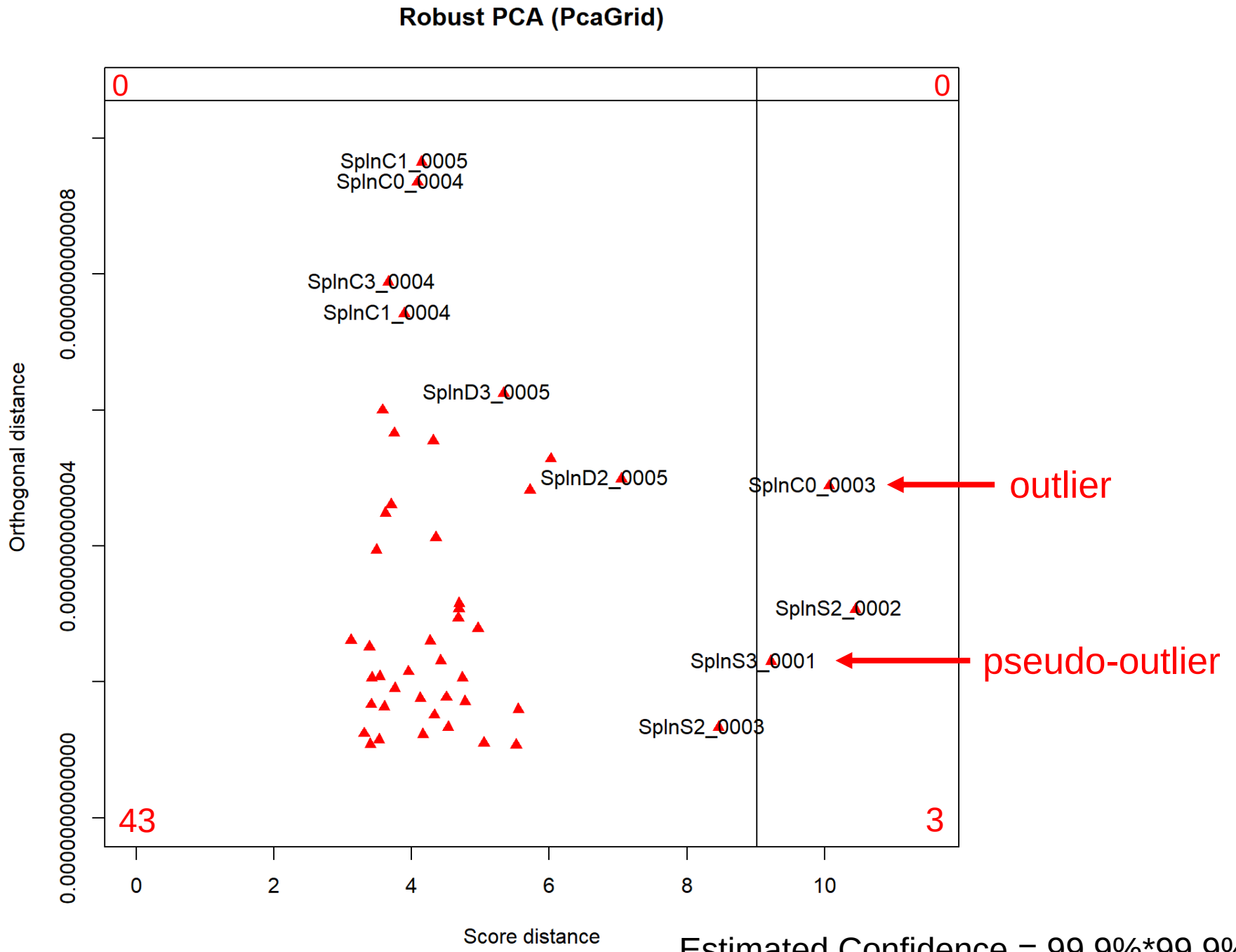

Empirical Histogram and Density versus the Standard Normal Distribution
$R^2 = 0.99$
Density
0.0
0.1
0.2
0.3
0.4
0.5
-6
-4
-2
0
2
4
6
Modified Z-Score

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

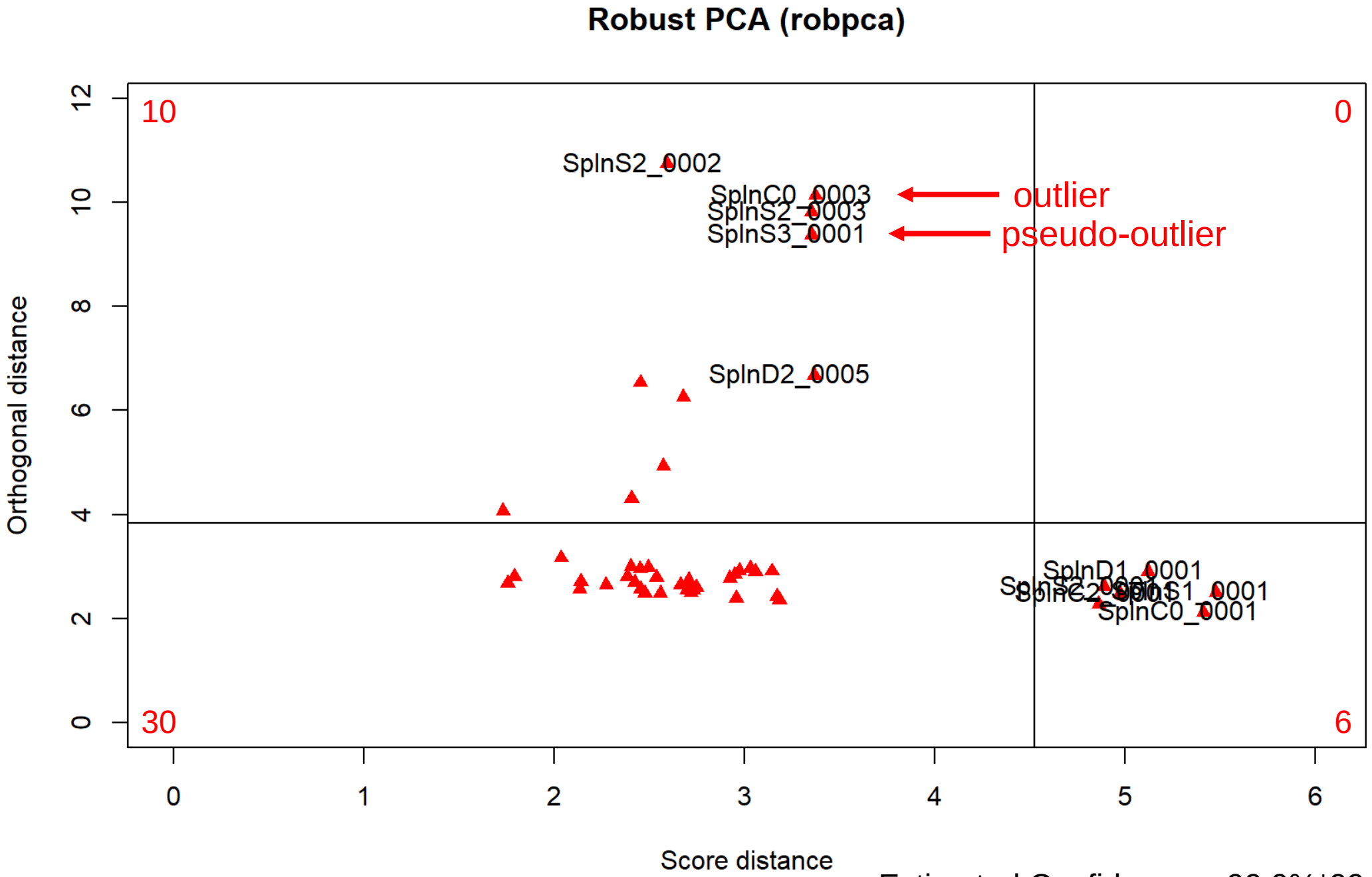


If SC < 0.25 (default parameter value), then the sample might be an outlier





Estimated Confidence = 99.9%*99.9% = 99.8%

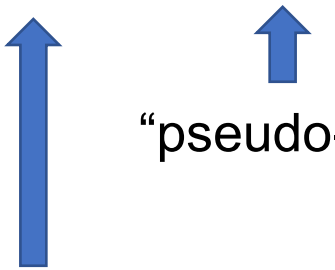


Estimated Confidence = 99.9%*99.9% = 99.8%

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] "SplnC0_0003" "SplnS3_0001"
```

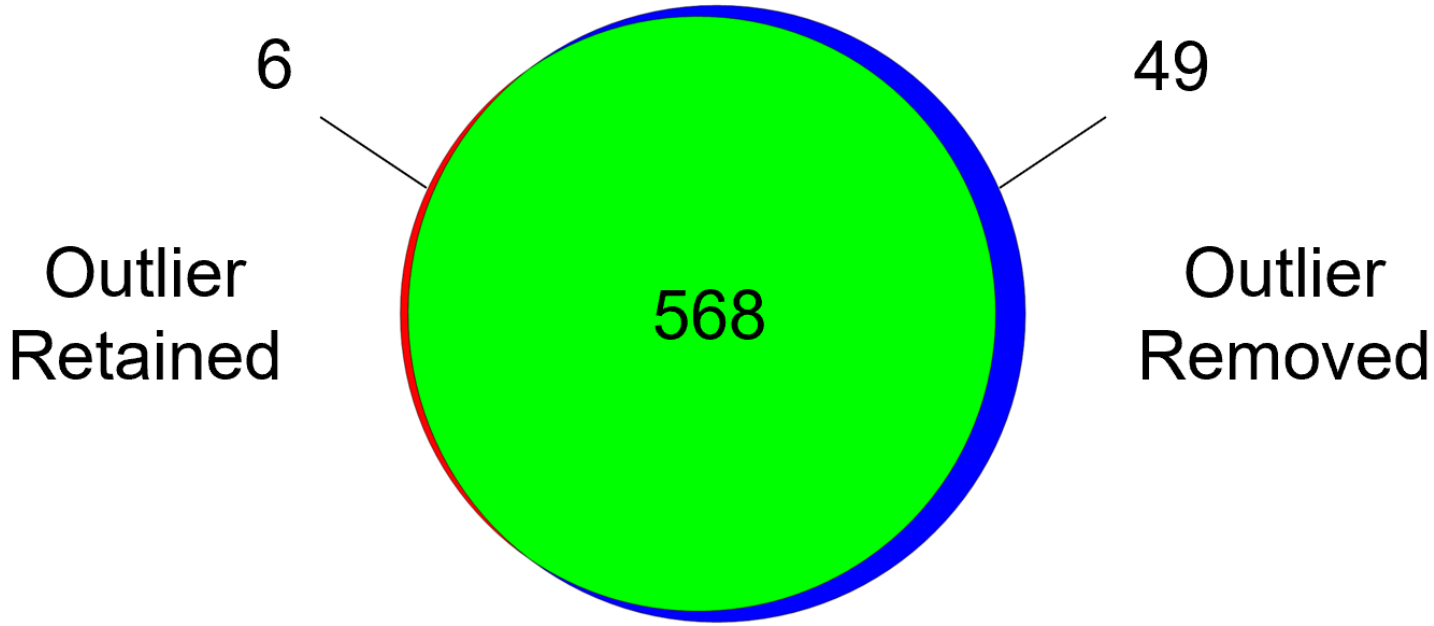


“pseudo-outlier”

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