Manual of the Standalone Version of NOREVA 2.0

February 11, 2020

TITLE: NOREVA 2.0: Enhanced Normalization and Evaluation of Time-course and Multi-class Metabolomic Data

DEVELOPER: Innovative Drug Research and Bioinformatics (IDRB) Lab <https://idrblab.org/>

MAINTAINER: Dr. Qingxia Yang <yangqx@cqu.edu.cn>

DESCRIPTION: The standalone NOREVA 2.0 enables the normalization of metabolomics data by: (**1**) realizing the normalization and evaluation of both *time-course* and *multi-class* metabolomic data, (**2**) integrating 144 normalization methods of a combination strategy, and (**3**) identifying the well-performing methods by comprehensively assessing the largest set of normalizations (168 in total). Because of the rapidly accumulating research interest in *time-course* and *multi-class* metabolomics, this study would make NOREVA unique in assessing normalization for this emerging field and could further enhance its popularity in metabolomics.

OPERATION PROCEDURE:

OPERATION PROCEDURES

1. Download and Install the Latest Release of *R* and *RStudio*

The latest release of ***R*** for Windows and its corresponding integrated development environment ***RStudio*** could be downloaded directly from (***1***) the NOREVA website:

https://idrblab.org/noreva2020/r-rstudio.zip

OR from (***2***) the official website of "*The R Project for Statistical Computing*" (https://cran.r-project.org/) and the official website of RStudio (https://www.rstudio.com/).

The whole installation should be completed by two sequential steps. First, please install the latest release of ***R*** by double clicking the executable file (**1-R-3.6.2-win.exe**) and following step-by-step instructions during the whole setup process (find more on https://www.r-project.org/about.html). Second, install the ***RStudio*** by double clicking the executable file (**2-RStudio-1.1.383.exe**) and following the step-by-step instructions during the whole setup process.

2. Download the Source Code of the Standalone Version of NOREVA

The source code of NOREVA together with the supporting ***R*** packages can be downloaded from:

https://idrblab.org/noreva2020/NOREVA.zip

Please decompress it by right clicking and selecting the "Extract to NOREVA\" in **Your Preferred Directory**. After running ***RStudio***, please change your working directory (in ***RStudio*** environment) to "**Your Preferred Directory**\NOREVA\NOREVA-Sourcecode\" by typing and then running the following ***R*** command:

```
setwd("Your Preferred Directory/NOREVA/NOREVA-Sourcecode/")
```

NOTE: (1) the ***R*** environment uses **forward slash** (/) to indicate the file path, which is different from the Windows CMD commands (backslash); (2) user can double check your current working directory by typing and then running the following ***R*** command:

```
getwd()
```

3. Run NOREVA *R* Scripts to Conduct Normalization and Assess Performance

3.1. Run NOREVA.R to normalizing and assessing the metabolomic dataset, which does not need to consider the internal standards (ISs)/quality control metabolites, using the following ***R*** command:

```
system(paste("Rscript  NOREVA.R", "Input-Dataset.csv", "-s", S , "-i", I, "-t", T))
```

When runing the above script, user should replace the **Input-Dataset.csv** to the name of their studied file. For example, there are six sample files in the working directory (in ***RStudio*** environment) “**Your Preferred Directory**\NOREVA\NOREVA-Sourcecode\”. Just copy the name (including the file extension .csv) to replace the **Input-Dataset.csv** in the above script. Moreover, before running the above script, **S**, **I** and **T** should be replaced by natural numbers. **S** is used to indicate the type of your input dataset, which can be 1, 2, 4, and 5. **I** is used to indicate the selected imputation method, and there are 4 selectable methods which is represented by 1, 2, 3, and 4. **T** is used to indicate the selected transformation method, and there are 3 selectable methods which is denoted by 1, 2, and 3. The detail information on these three arguments can be found in **Table 1** below.

The required formats of input data are provided in those four sample files in the working directory (in ***RStudio*** environment) “**Your Preferred Directory**\NOREVA\NOREVA-Sourcecode\”, which include: ‘Sample.Data-Multiclass-QCS.csv’, ‘Sample.Data-Multiclass-NONE.csv’, ‘Sample.Data-Timecourse-QCS.csv’, and ‘Sample.Data-Timecourse-NONE.csv’.

3.2. Run NOREVA.R to normalizing and assessing the metabolomic dataset, which consider internal standards (ISs)/quality control metabolites, using the following ***R*** command:

```
system(paste("Rscript  NOREVA.R", "Input-Dataset.csv", "-s", S, "-i", I, "-t", T, "-is", "IS"))
```

When runing the above script, user should replace the **Input-Dataset.csv** to the name of their studied file. Just copy the name (including the file extension .csv) to replace the **Input-Dataset.csv** in above script. Moreover, before running the above script, **S**, **I** and **T** should be replaced by natural numbers. **S** is used to indicate the type of your input dataset, which can be 3 and 5 in this situation. **I** denote the selected imputation method, and there are 4 selectable methods which is represented by 1, 2, 3, and 4. **T** is used to indicate the selected transformation method, and there are 3 selectable methods which is denoted by 1, 2, and 3. **IS** should be a series of natural numbers separated by comma. For example, the replacement of **IS** to 2,6,9,n indicates that the metabolites in the 2st, 6th, 9th, and nth columns of in your input dataset **Input-Dataset.csv** should be considered as the ISs or quality control metabolites. The detail information on these four arguments can be found in **Table 1** below.

The required formats of input data are provided in those two sample files in the working directory (in ***RStudio*** environment) “**Your Preferred Directory**\NOREVA\NOREVA-Sourcecode\”, which include: ‘Sample.Data-Multiclass-IS.csv’ and ‘Sample.Data-Timecourse-IS.csv’.

Table 1. Description of the arguments in the ***R*** command.

Argument Name	Description and Utility of the Corresponding Argument
Input-Dataset.csv	**The Name of the Input File (txt or csv format)** Please find the detail information of the file format from those six sample files in the working directory (in ***RStudio*** environment) "**Your Preferred Directory**\NOREVA\NOREVA-Sourcecode\"
-s	**Flag of the Selection of Style of dataset** If set **1**, the multi-class dataset without QC samples and ISs If set **2**, the multi-class dataset with QC samples If set **3**, the multi-class dataset with ISs If set **4**, the time-course dataset without QC samples and ISs If set **5**, the time-course dataset with QC samples If set **6**, the time-course dataset with ISs The default setting of this flag is "**1**"
-i	**Flag of the Selection of Imputation Methods** If set **1**, method of column mean imputation If set **2**, method of column median imputation If set **3**, method of half of the minimum positive value If set **4**, method of KNN imputation The default setting of this flag is "**1**"
-t	**Flag of the Selection of Transformation Methods** If set **1**, method of cube root transformation If set **2**, method of log transformation If set **3**, none transformation method The default setting of this flag is "**1**"
-is	**Flag of the Column of Internal Standards (IS)** If there is only one IS, the column number of this IS should be listed If there are multiple ISs, the column number of all ISs should be listed and separated by comma (,)

This command enables the performance assessment of all 168 normalization methods based on 4 distinct criteria. These six datasets (csv files) include: (**a**) multi-class dataset without quality control samples (QCS) and internal standards (ISs) named by Sample.Data-Multiclass-NONE.csv; (**b**) multi-class dataset with QCS named by Sample.Data-Multiclass-QCS.csv; (**c**) multi-class dataset with ISs named by Sample.Data-Multiclass-IS.csv; (**d**) time-course dataset without QCS and ISs named by Sample.Data-Timecourse-NONE.csv; (**e**) time-course dataset with QCS named by Sample.Data-Timecourse-QCS.csv; (**f**) time-course dataset with ISs named by Sample.Data-Timecourse-IS.csv.

After running the above ***R*** command, a variety of OUTPUT files are generated as following:

Name	**Output Type**	**Description**
OUTPUT-NOREVA-Overall.Ranking.Figure.pdf	PDF Figure	A heatmap illustrating the performance ranking of all normalizations based on the criteria selected by user
OUTPUT-NOREVA-Overall.Ranking.Data.csv	CSV File	A CSV file containing all information of performance assessment, criteria selected and ranking
OUTPUT-NOREVA-All.Normalized.Data.Rdata	RDATA File	A RDATA file providing the resulting outcomes of all normalization methods
OUTPUT-NOREVA-Criteria.Ca	Folder	A number of PDF files illustrating the PMAD plot (intensities among replicates) of each normalization
OUTPUT-NOREVA-Criteria.Ca.Rdata	RDATA File	A RDATA file providing the PMAD value (intensities among replicates) of each normalization
OUTPUT-NOREVA-Criteria.Cb	Folder	A number of PDF files illustrating the *k*-means clustering (distinct groups) of each normalization
OUTPUT-NOREVA-Criteria.Cb.Rdata	RDATA File	A RDATA file providing the purity of *k*-means cluster ability on distinct groups of all normalizations
OUTPUT-NOREVA-Criteria.Cc	Folder	A number of PDF files illustrating Venn diagram for the marker overlap of each normalization
OUTPUT-NOREVA-Criteria.Cc.Rdata	RDATA File	A RDATA file providing the *CWrel* value of marker overlap of all normalizations
OUTPUT-NOREVA-Criteria.Cd	Folder	A number of PDF files demonstrating the marker classification of each normalization
OUTPUT-NOREVA-Criteria.Cd.Rdata	RDATA File	A RDATA file providing the AUC value for marker classification of all normalizations