### EnsMOD Monte Carlo (EnsMOD MC) 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@list.nih.gov

#### Introduction

If one or more outliers are detected using EnsMOD, EnsMOD\_MC can be used to reanalyze the user's dataset to estimate the outlier detection false positive rate (FPR, which equals the probability of falsely detecting an outlier). EnsMOD\_MC uses the Monte Carlo method, and a good introduction to the Monte Carlo method is William L. Dunn and J. Kenneth Shultis, Exploring Monte Carlo Methods 2nd Ed., Elsevier Publishing, 2022.

EnsMOD\_MC does not make any assumptions about the user's dataset (e.g., approximately Gaussian variance is not presumed). EnsMOD\_MC is available as a stand-alone Rmarkdown script. EnsMOD and EnsMOD\_MC are open-source and freely available (https://github.com/niaid/EnsMOD).

#### **EnsMOD MC Overview**

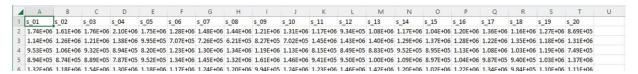
First, delete the detected outlier(s) from the input dataset and save it as "Gene\_Expression\_Table.xlsx". The resulting number of columns needs to be ≥9. Run EnsMOD\_MC. Use the same four cutoff parameter values (CCC, SC, robpca, PcaGrid) that were used to detect the outlier(s). One row (e.g., gene) at a time, some or all of the input data are shuffled. The EnsMOD HCA and rPCA tests are performed, and false positive outliers are detected, if any. This is repeated a preset number of times. The estimated FPR is the number of detected outliers (i.e., false positives) divided by the number of simulations.

By default, all columns are included during shuffling. To prevent a column from being shuffled, append "\_fixed" to the end of the column header (i.e., the sample unique identifier) of the input dataset. The experimental conditions that did not correspond to an outlier do not need to be shuffled. If multiple experimental conditions correspond to detected outliers, they can be analyzed separately using EnsMOD\_MC.

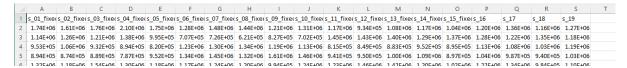
## **Installing and Running EnsMOD\_MC**

- 1. Install R (<a href="https://www.r-project.org/">https://www.r-project.org/</a>).
- 2. Install RStudio (https://www.rstudio.com/).
- 3. Acquire "EnsMOD\_Monte\_Carlo\_stand-alone\_v1\_0.Rmd" (https://github.com/niaid/EnsMOD).
- 4. The script automatically installs and updates all of the required R packages.
- 5. Provide the table of input data.
  - a. The input table needs to be in "Gene\_Expression\_Table.xlsx", and it needs to be located in the same directory as the EnsMOD\_MC Rmarkdown file.
  - b. The XLSX file should contain only one worksheet and only one table.
    - i. The first row (the header) should contain unique identifiers of the samples.
    - ii. The detected outlier columns should be deleted from the table.
    - iii. To prevent a column from being shuffled, append "\_fixed" to the end of the column header (i.e., the sample unique identifier).
  - c. 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).
  - d. EnsMOD example datasets are provided at (https://github.com/niaid/EnsMOD/tree/main/app/EnsMOD\_Examples)
  - e. During our testing of EnsMOD, a minimum of nine samples (i.e., columns) were required. Fewer samples resulted in the robpca step failing ("Error in robpca: Something went wrong with the outlyingness computations.").

Original simulated proteomics dataset used for EnsMOD (s 20 is the simulated outlier).



Simulated proteomics dataset used for EnsMOD\_MC. Sample s\_20 was deleted, and only the fourth experimental condition (it was s\_16, s\_17, s\_18, s\_19, s\_20) was shuffled.



# 6. Ensure that the EnsMOD\_MC settings are correct.

- a. Use RStudio to edit the parameter values and click Save.
- b. EnsMOD and EnsMOD\_MC should use the same four outlier detection cutoff parameter values. They are:
  - i. The minimum CCC threshold
  - ii. The maximum SC threshold
  - iii. The robpca probabilistic threshold
  - iv. The PcaGrid probabilistic threshold
- c. The number of Monte Carlo simulations (num\_MC\_Tests) should be at least ~500, but note that the runtime could be many hours or even days (try 10 to estimate the runtime for 500 simulations).

## 7. Run EnsMOD\_MC.

a. To run EnsMOD\_MC, click the "Knit" button. EnsMOD\_MC will produce an HTML output file that can be reviewed using a web browser.

