Harman tutorial

Harman package is used for batch effect removal in Micro Array Gene Expression (MAGE) datasets using Harman method. It is based on batch-mean shrinkage for every individual principal component. The highest possible shrinkage is implemented for each principal component, with the constraint that the confidence in not overcorrecting the data which means not losing genuine biological data,is above the threshold set by the user (eg: 95%). This confidence threshold is related to (i.e the complement of) the statistical p-value.

Harman is developed under Matlab2013a environment and to run its executable file(Harman.exe) you should first install MCR(Matlab Compiler Runtime) 8.1 from this link(www.......)on your machine. Having installed MCR 8.1, double clicking on Harman.exe starts the execution of the program and the input GUI would be open as shown in Fig1. This process may take up to 30 sec on some computers.

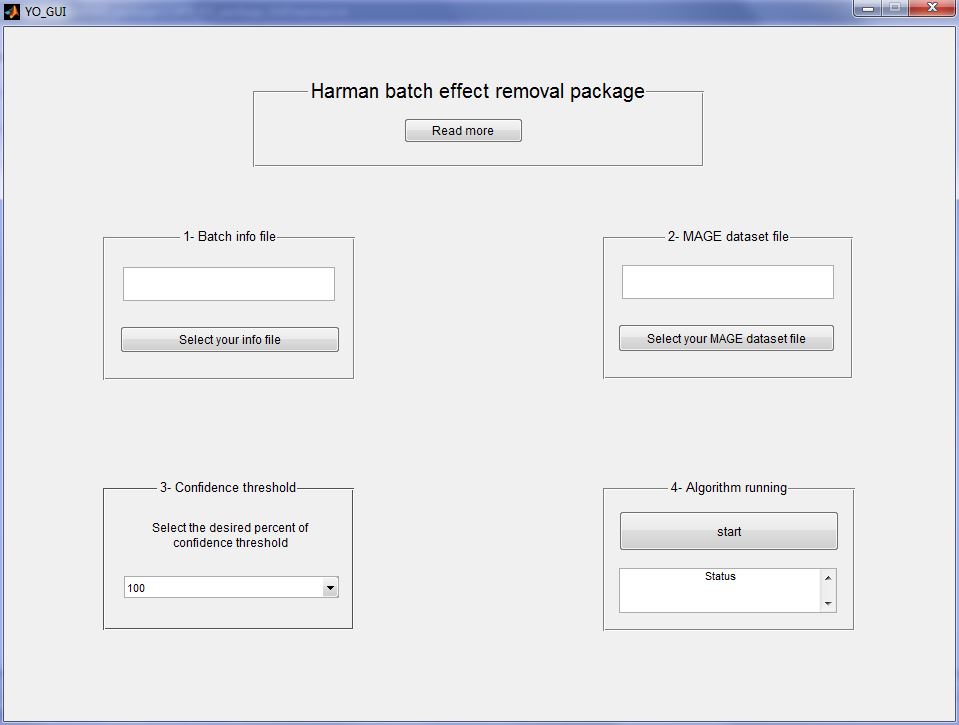


Figure 1: Harman Input GUI

All you need to use this package are two data and information files which as an example are shown in Figure2.

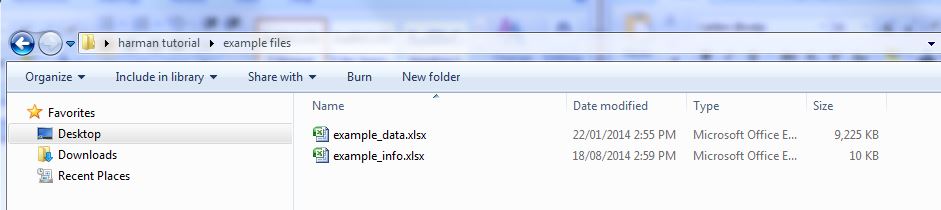


Figure 2: Two example files

The example data file is a (33298\*29)=(p+1) by (n+1) tabular dataset in excel format(xlsx) with p=33297 probesets as the rows and n=28 replicates arrays as the columns. The column headings correspond to arrays name and the row headings correspond to probeset IDs as shown in Figure 3.

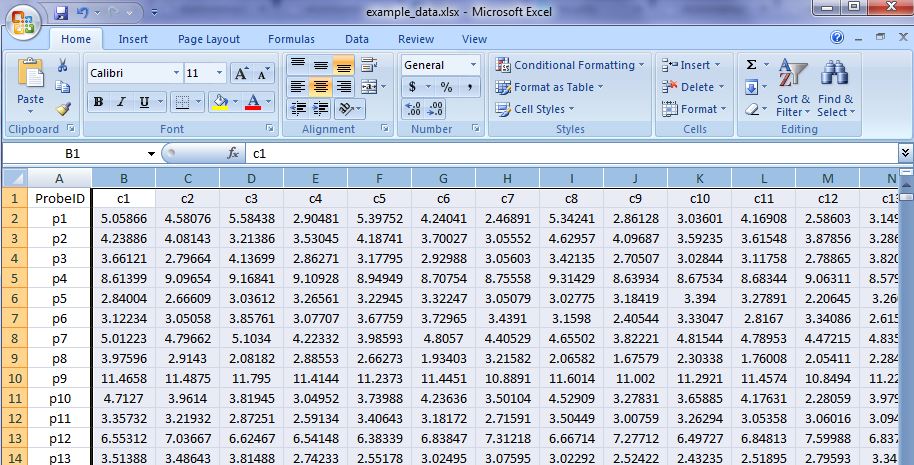


Figure 3: Sample data files

The example info file contains information about the treatment types and the batch number of each replicates arrays. Then as shown in Figure 4 the column headings of the data file is the same as the row headings of the info file.

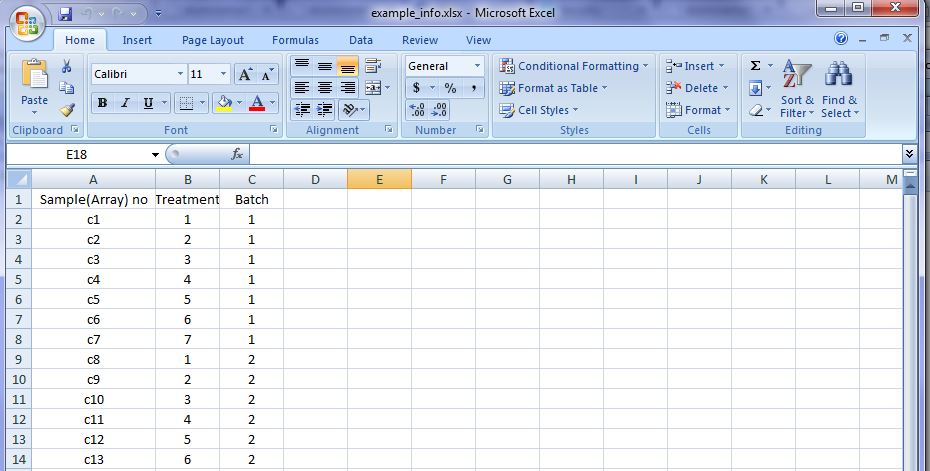


Figure 4: Sample info file

To feed the package with these two files, we start two consecutive file selection processes by pressing “Select your info file” and “Select your MAGE dataset file” buttons .The selected files are displayed in the provided boxes as shown in Figure 5,6,7,8.

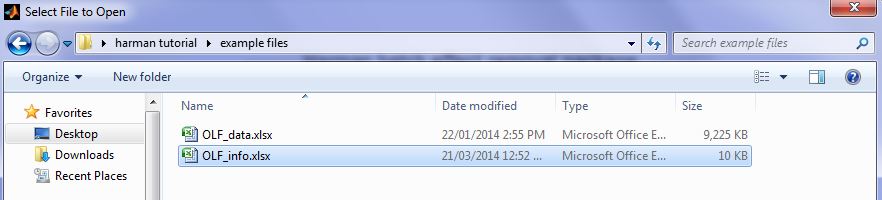


Figure 5: Info file selection

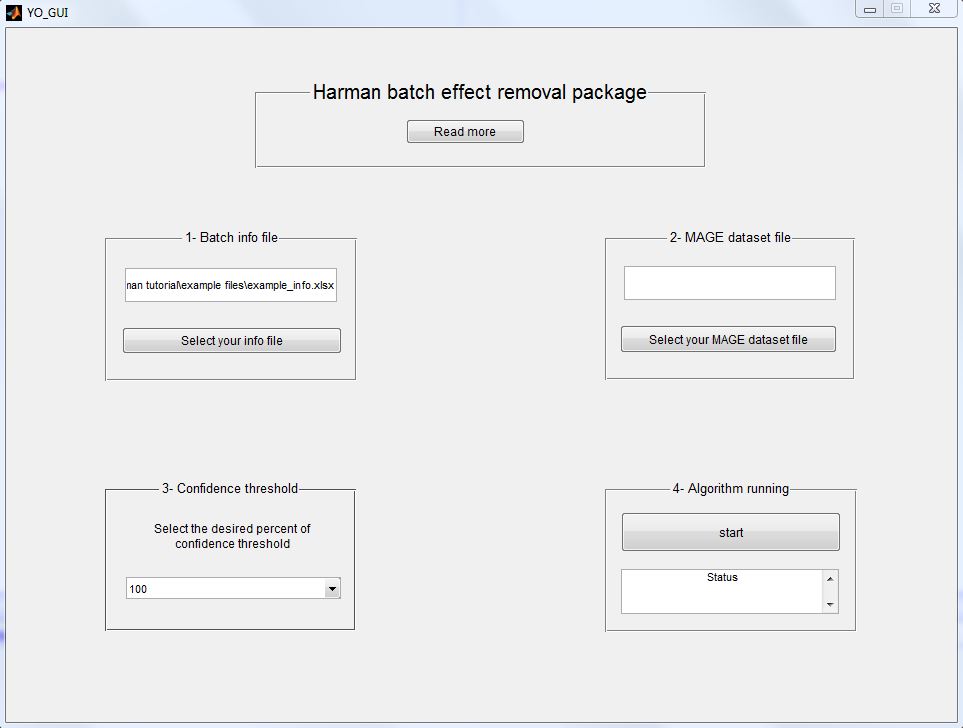


Figure 6: Info file selected

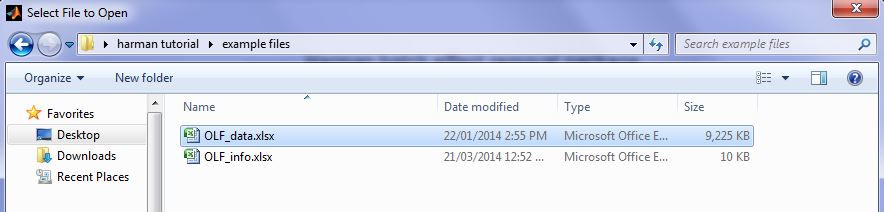


Figure 7: Data file selection

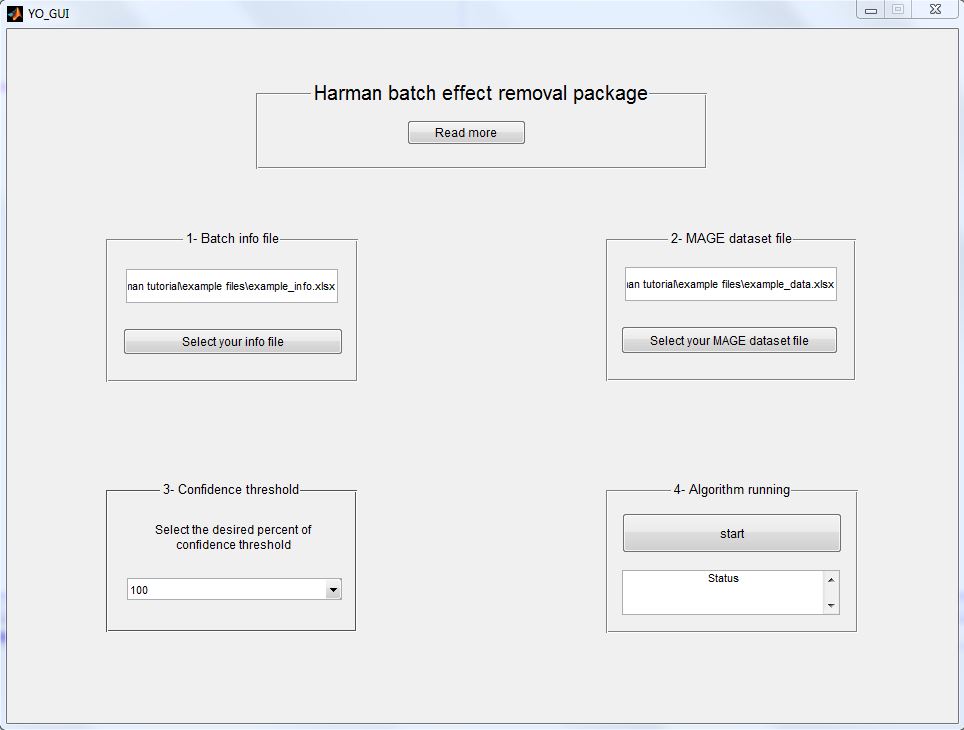


Figure 8: Data file selected

Having fed these files, the user should select the confidence threshold from the popup menu as shown in Figure 9. This number shows our confidence in not overcorrecting the data and losing genuine biological information. The most common used value is 95%.

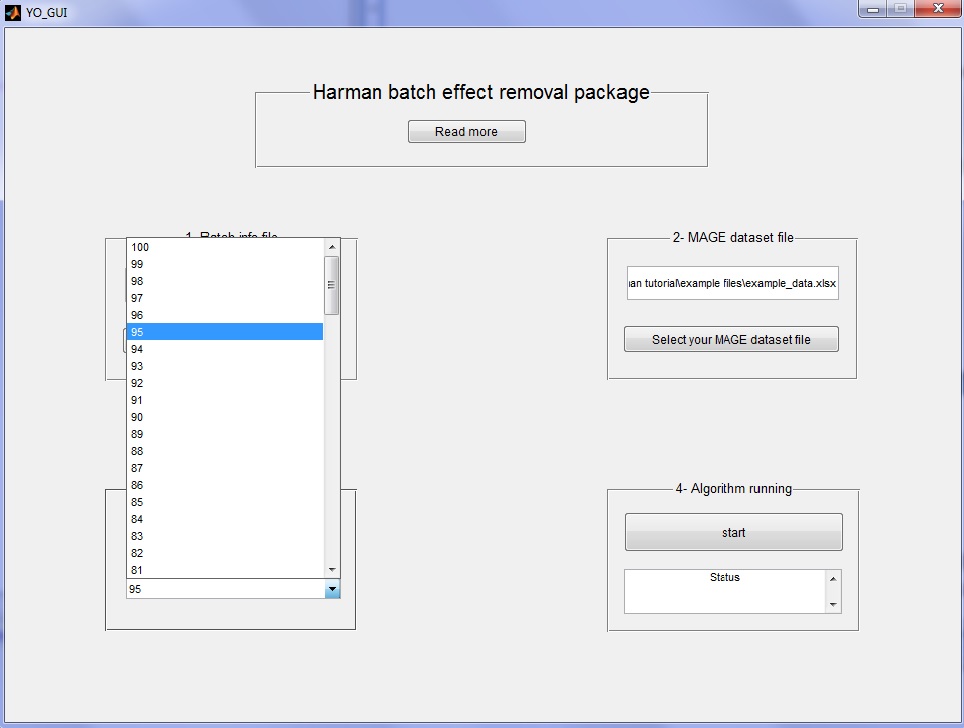


Figure 9: Selecting the confidence threshold

Now, everything is ready to run the algorithm. By pressing the start button, the package would start running and a progress bar would dispaly the progress of the algorithm as shown in Figure 10. It should be mentioned that depending on the batchiness severity and the number of samples in dataset, it may take a couple of minutes for the pop up progress bar to be updated in some occasions.

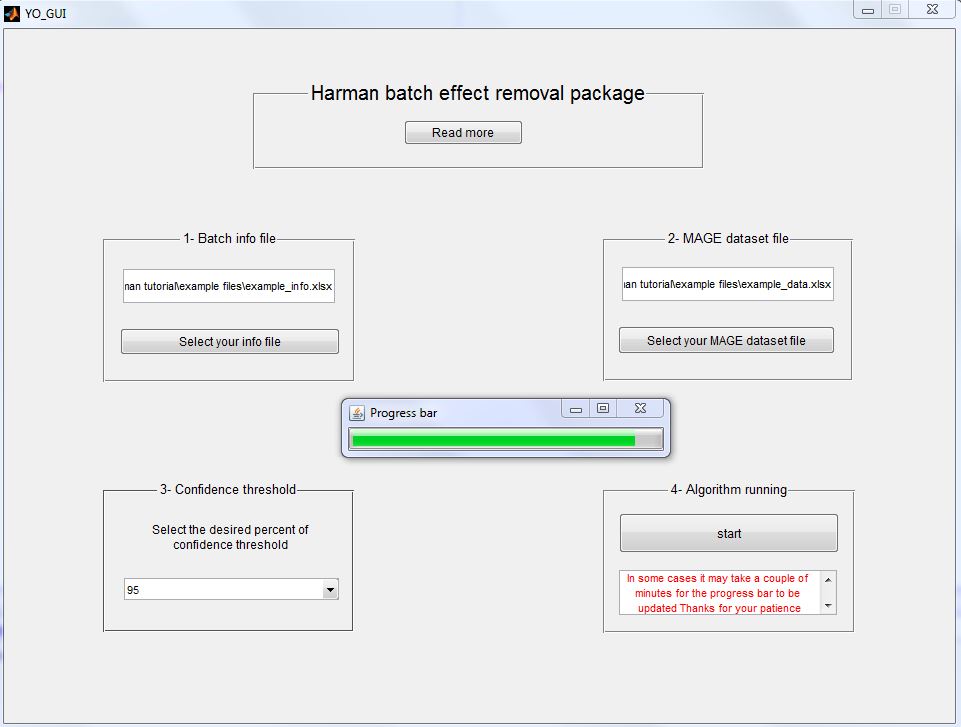


Figure 10: Running progress bar

Having the algorithm running finished, another PC visualization tool GUI would pop up which shows the overall correction vector for all principle components along with the two plots of selective principle components before and after batch effect correction, Figure 11.

At the same time the corrected MAGE dataset would be saved in the same folder as the example data with a –HarmancorrectedXX.xlsx postfix where XX shows the chosen confidence threshold. The correction vector and the raw and corrected PCs are also saved in two excel files with –correctionvectorXX.xlsx and \_PCrawcorrXX.xlsx postfixes.

The effectiveness of the batch correction algorithm can be visually explored from the PC plots. For the example dataset, the two highest corrections correspond to two smallest correction numbers, 0.25 and 0.33, which belongs to PC1 and PC2 respectively. If we choose PC1 and PC2 as PC-X and PC\_Y, then as shown in Figure 12, we see the effect of batch noise which moves batch 3 scatters(Cyan scatters) to a separate cluster from other batches scatters in the left plot. By removing the batch noise, this cyan cluster has moved toward other batches scatters in the right plot which reflects the performance of the batch effect correction algorithm

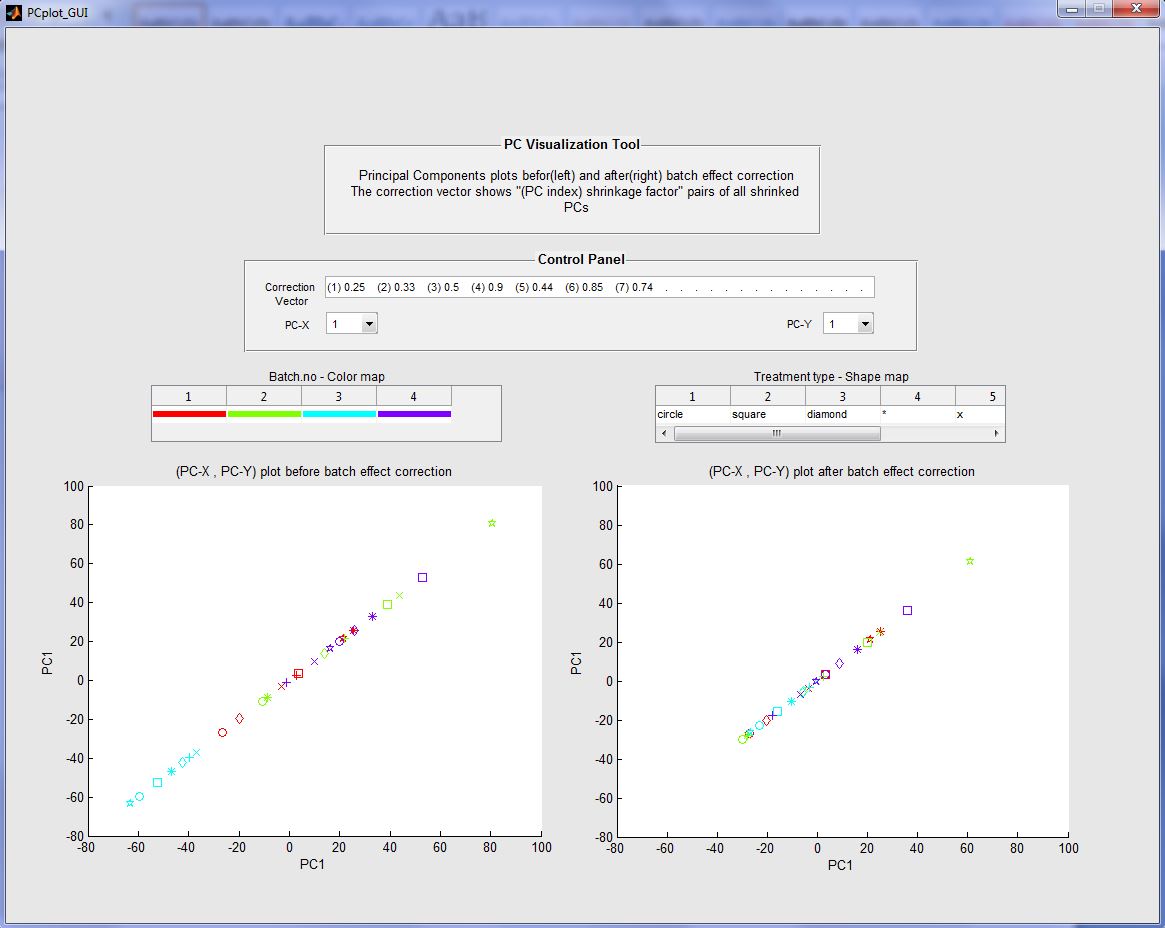


Figure 11: PC visualization tool GUI