

Direction for using the R code "functional category analysis.txt" to obtain p-values and plots for the identification of over-represented or differentially expressed functional gene category:

Step 1. Directory:

- Choose a directory that you want to work with for this analysis. For example, "C:/fca". This directory should contain the two data sets as specified in step 2. All the R output will be exported into this directory.

Step 2. Data format:

- Population data: an Excel spreadsheet that contains two columns. The first column contains all probesets in the microarray experiment; the second column contains corresponding p-values that were obtained from previous statistical tests for differential gene expression, for example, the p-value from a t-test. This file should NOT contain a header row. In Excel, save this file as a .csv file in the directory you chose in step 1, for example, "c:/fca/population.csv". An example of this data set, "population.csv", is provided in the Additional File 3.
- Functional gene category data: an Excel spreadsheet that contains one column. The column contains all probesets in a specific functional gene category. This file should NOT contain a header row. In Excel, save this file as a .csv file in the same directory, for example, "c:/fca/apoptosis.csv". An example of this data set, "apoptosis.csv", is provided in the Additional File 4.

Step 3. "functional category analysis.txt":

- Open the file "functional category analysis.txt".
- Replace "type your directory here" under (1) by the directory that you chose in step 1. For example, the code can be changed to: `setwd("c:/fca")`.
- Replace "type the name of your population data here" under (2) by the name of your population data that you have in step 2. For example, the code can be changed to: `population=read.csv("population.csv",header = FALSE)`.
- Replace "type the name of your functional category data here" under (2) by the name of your functional gene category data that you have in step 2. For example, the code can be changed to: `fc=as.matrix(read.csv("apoptosis.csv",header = FALSE))`.
- Open R. Run "functional category analysis.txt" by copying all code in it and paste into R.
- Optional: note that the alpha level used by Fisher's exact test and EASE analysis is set as 0.05 in this code. Users can change the alpha level by replacing "0.05" under (3) by a different alpha level. For example, the code can be changed to: `alpha=0.01`. The user then need to copy the entire code and paste into R.

Step 4. Results:

- Go to the working directory that you specified in steps 1 and 3, you will see 4 files that are generated by R, which are:

1. "FCA pvalues.csv": contains EASE score, and p-values from Fisher's exact test, Wilcoxon Mann -Whitney test, and two-sample and one-sample Kolmogorov-Smirnov tests.
2. "FCA-ECDF plot.jpg": contains the ECDF plot.
3. "FCA-ECDF plot enlargement.jpg": contains the enlargement of the lower left corner of the ECDF plot.
4. "FCA-ROC curve plot.jpg": the ROC curve plot.

Notice that you might see the following warning message in the R window after you paste code in "functional category analysis.txt" into R:

Warning message:

1: cannot compute correct p-values with ties in: `ks.test(fc.p, other.p, alternative = "gr")`

2: cannot compute correct p-values with ties in: `ks.test(fc.p, "punif", alternative = "gr")`

The reason for you receiving this warning message is because that you have at least two probesets in your input data set that have the same p-values. In this case, R will return an approximate p-value rather than an exact p-value for the Kolmogorov-Smirnov test. However, this approximation is OK because of the large samples in the functional category analysis (a functional gene category usually contains several hundreds of probesets). Thus the user can ignore this warning message.