README: 200415 Syllogist: Cell type deconvolution using RNA-seq data (FISHER TEST)

This is an R script to be preferentially run within RStudio

Versions used to run the code:

R version: 3.6.1
RStudio version: 1.2.5001

BEFORE START

1) Required packages to run the script:

a.BiocGenerics v0.32.0 (Bioconductor)
b.pheatmap v1.0.12 (CRAN)
c.gplots v3.0.1.1 (CRAN)
d.viridis v0.5.1 (CRAN)
e.RColorBrewer v1.1-2 (CRAN)

2) Required files to run the script "SYLLOGIST_Rscript.R"
(all files are provided with the script and need to be placed in the same working directory)

a. The Reference Map ("200116_Reference map_Top 80 Specific genes for each cell type.csv"). This data is also provided as Supplementary Table 1 in the manuscript

b. A list of gene names from which the script will perform random sampling (provided as "Affy_genes_unique3.csv")

c. The query data to be separated in cell types in .csv format (we provide 1 example dataset:
"Example query 1_GSE59612_GBM cores vs normal brain tissues.csv" from Gill et al. PNAS. This dataset has been used to generate results of Figure 2 a,b and supplementary Figure 2a of the manuscript.
d. OPTIONAL: To run your own query data, please provide a file with transcriptome data in .csv format with gene names (HUGO Symbols) as rows and samples as columns (see example provided for more details).

3) Required action: set working directory by updating the setwd() function accordingly at line 54 of the script

4) script can be ideally run using the source button in RStudio

CODE EXECUTION

1) As soon as all required packages are installed, all required files are placed in the working directory, and the setwd() line is updated accordingly, you can source the script via RStudio

2) A list of files contained in the working directory will be shown in the console and the source area of RStudio

3) A first prompt asking "which data should I analyze ? --" will be shown in the console. Enter the number of the file containing the query dataset " Example query 1_GSE59612_GBM cores vs normal brain tissues_FPKM" as seen in the list (e.g. "**3**")

4) A second prompt will ask "Choose a threshold". This threshold is defined to distinguish between expressed genes and background noise in the transcriptome data and varies depending on the expression data used. The query transcriptome provided is in FPKM units, therefore, one common threshold used is **0.1** (the same threshold used in the manuscript).

5) A third prompt will ask "How many Monte Carlo simulations?" for the number of simulations desired to run the null model. In the manuscript we used **1000** for all analyses. By using less simulations (e.g. **100**), the script will run much faster but with slightly less precision.

RESULTS

1) Result data are stored in object "**Fisher.t**", which contains normalized odds ratios as proxy for the relative quantity of the respective cell type. This object contains normalized odds ratios. **IMPORTANT**: the resulting odds ratios can be used to compare the same cell type between samples, but cannot be used to compare different cell types within one sample.

2) Raw odds ratios are also stored in object "**Fisher.t_unscaled"**

3) Heatmap is generated from the Fisher.t file to visualize data and shown in the plots