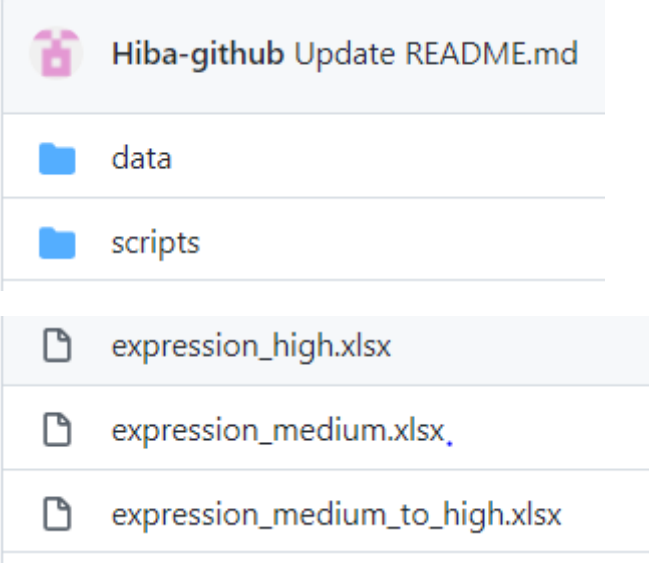
Reproducibility guide

Part 1. Script adaptation for new input files , For MS* disease data :

All processed not filtered data are provided as supplementary files , so the user can define different settings.

Option 1 : Change the p value to define Differentially expressed genes

Note : the p value used is <0.05



	A	B	C
1	Gene.symbol	P.Value	logFC
2	CYP11B2	0.000327	1.01e-01
3	MACC1	0.000517	-2.49e-01
4	F13A1	0.000752	2.95e-01

Step 1 :

Access the ''data'' directory.

Download the target prevalence

Expression Excel file

Step2 :

Open the file in Excel and filter the column P.Value according to your settings.

Step3 :

Save the file (same file name)

Option 2 : Change the target genome region for CpGs methylation

Note : In the case study we targeted the promoter region.

Supplementary files are provided for "gene associated cpgs " and "non gene associated cpgs"

- gene_associated_methylation_high.xlsx
- gene_associated_methylation_low.xlsx
- gene_associated_methylation_medium.xlsx
- gene_associated_methylation_medium_to_high.xlsx
- nongene_associated_methylation_high.xlsx
- nongene_associated_methylation_low.xlsx
- nongene_associated_methylation_medium.xlsx
- nongene_associated_methylation_medium_to_high.xlsx
- promoter_methylation_high.xlsx
- promoter_methylation_low.xlsx
- promoter_methylation_medium.xlsx
- promoter_methylation_medium_to_high.xlsx

Step 1 :

Access the ''data'' directory.

Download the target region (Gene associated or non gene associated)

Expression Excel file

Step 2 :

In the script replace the file nane.

```
# Is gene promoter methylated in medium to high prevalence countries --------



#methylation profile study
#replace the gene symbol and Run lines
promoter_methylation_medium_to_high<- read_excel("promoter_methylation_medium_to_high.xlsx")
cpg_met2 <- promoter_methylation_medium_to_high %>%
  filter(gene =="STAT3")
```

Part 2 : Reproducibility for other datasets

1. Getting data from GEO database:

This workflow is applicable for publically available data on the NCBI GEO database.
On the GEO website we searched for the keyword “Multiple Sclerosis” and the species filter was set to “Homo Sapiens”. And only experiments that used blood samples were included.
The datasets were chosen as following :

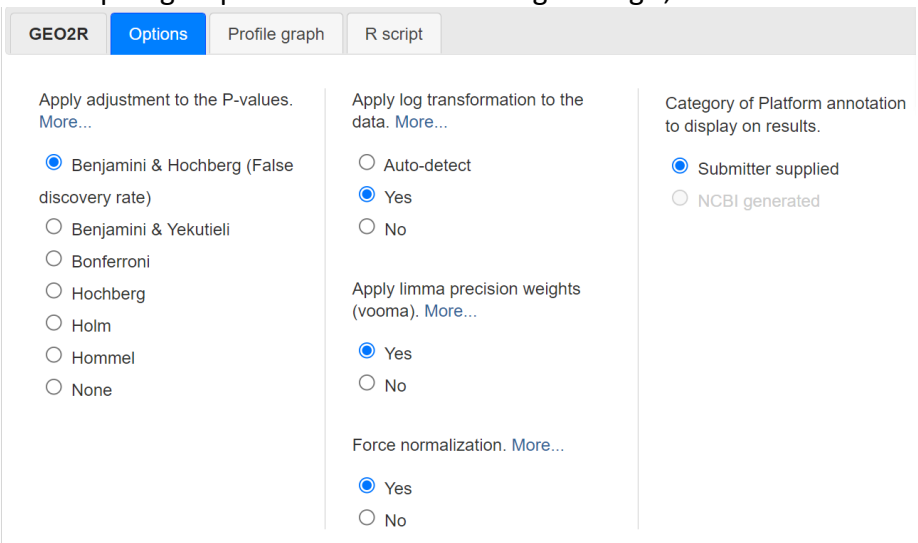
2. Genetic expression data :

Use the study type filter: all “Expression profiling “ in GEO database

- If the dataset data is FastQ files, use SRA database to download the data.

Use FASTQC tool for quality control, trim the data with Trimmomatics tool, next Alignment to the genome using Hisat Software .Finally , last use R studio for RNA quantification.

- If dataset data are txt count files, use EdgeR R package to determine differencial expression data.
- If dataset data are Cel files , use limma package for normalization , annotate the data , and analyze differencial epression with DESeq2 .
- For same datasets the GEOR tool is availbe to do online expression analysis: define the sample sgroups then use the following settings ;



Convert all the final expression data files to .xlsx files .

You only need 3 column :
Column 1 : name =”gene” ; content = gene symbols
Column 2 : p value
Column 3 : LogFc

To identify the differencially expressed gene:

- use the filter function of Excel to only keep the genes with $p<0.05$
- use the filter function again to oder the LogFc
- divise the data on 2 excel files :
  - one only containing the genes with positive logFC value , only keep the gene column and save the file as csv “ up_exp.csv”

- -one only containing the genes with negative logFC value , only keep the gene column and save the file as csv " down_exp.csv"

Note : For the genetic profiling data , multiple analyzing methods were used and mentioned, for different data format on GEO database :

- Analysis of FastQ files from SRA database

- Analysis of Count txt files

- Analysis of CEL files

- And GEO2R tool for big size data.

3. Methylation data :

Use the study type filter: all "Methylation profiling " in GEO database
For the methylation profiling data, idat files are the most common raw data format:

Import Data (idat files and phenotype data) in R studio using GEOquery R package
Use minfi R package for for data processing the annotate the data.
Use Limma R package to identify differentially methylated CpGs.
Export that final data as .xlsx file .
Name the column containing gene symbols : "gene "
Filter the data using excel filtering function , based on the genomic region (gene associated, non gene associated or promoter) .

4. MicroRNA expression study :

Use The study type filter: all "Non-coding RNA profiling " were applied for microRNA data.

Only GEO2R tool was used to analyze the miRNA in this study .
Export result data as xlsx files (colums : "miRNA ID", 'gene","p values" and "log FC value")

5. Adaptation to the R script:

In order to be able to use the same R Script , make sure to name the final data files exactly as the MS case study files are named or just change the commands lines to import the files.