Team Meeting

08.04. 2022 / 2:00 PM / Zoom

Attendees

Ate, Xavier, Ni, Stefan, Zexi

Agenda

Introduce the responsibilities of two groups:

Group1: Zexi & Ate
Group2: Xavier & Ni & Stephan

Library used

library("GEOquery") - the name of the dataset is 'GSE105018_family.soft'

library("illuminaio")

Definition and understanding on GPL, GSM, GSE

GSE: the task we are focusing (Whole blood DNA methylation profiles in participants of the Environmental Risk (E-Risk) Longitudinal Twin Study at age 18.)
GSM: samples (1658)
GPL: platform of the data (GPL1353 Illumina HumanMethylation450 BeadChip)
The information about 450k variables (CPG position)

Data source 1: GSE105018_family.soft - GPL (not important)

Code: data = getGEO(GEO = 'GSE', filename = "GSE105018_family.soft", GSEMatrix = TRUE)
Path: data$gpls$GPL13534$dataTable$table (485577 * 37)

Data source 2: GSE105018_series_matrix.txt.gz - GSM

Code: gse <- getGEO("GSE105018", GSEMatrix = TRUE)
Path: gse$GSE105018_series_matrix.txt.gz$phenoData$data (1658 * 40)

Data source 3: GSE105018_RAW.tar (

Code: readIDAT("")

Applied Methods for Analysis:

- Discovery Analysis: They did some analysis on the main dataset (NTR)

- Replication Analysis: To determine if the basic findings of the original study can be applied to other participants and circumstances, the repetition of the study is done on different datasets (E-Risk, FTC, TwinsUK, BSGS, NTR Children)
- Sensitivity Analysis: Sensitivity analysis was performed to examine the robustness of the findings. In this paper, sensitivity analysis was conducted in the NTR and BSGS datasets, because these cohorts also had DNA methylation data available for non-twins.

Note:
The IDAT file can be obtained from either the 13GB file or the two “supplementary file” columns in 1658 data.

Each sample has two corresponding supplementary files (green & red), each containing 450k rows, each containing the SNP value (our predictor?). Then we use this data / feature obtained from this data to merge with our sample data then test the classifier.

Interpretation: Using 450k * 2 variables to predict 1658 samples.

AddressA_ID is contained in both the IDAT file and our 1658 data, so can be used as a primary key.

To Do List

1. Exam paper to see how they build the classifier
2. Keep exam the first part of the paper(team1)
3. Ask for the client's help if the 1st step for team2 has not been progressing.