**Exposome Data Challenge – Team Brunius**

Anton Ribbenstedt (anton.ribbenstedt@chalmers.se)  
Elise Nordin (elise.nordin@chalmers.se)  
Tessa Schillemans (tessa.schillemans@ki.se)  
Carl Brunius ([carl.brunius@chalmers.se](mailto:carl.brunius@chalmers.se))

**Overview**

This data analytical pipeline is divided into 3 main parts:

1. *Exposome-Phenotype Modules* to identify exposures of interest
2. *Exposome-Omic Modules* to identify Omics variables of interest
3. *Exposome-Omic-Outcome Modules* to combine exposures, omics and outcomes of interest

In the following descriptions, MUVR refers to our in-house algorithm for machine learning modelling withing a repeated double cross-validation with minimally biased variable selection (Shi et al 2019. Bioinformatics)

1. ***Exposome-Phenotype Modules***

Steps:

1. The first step was to convert all the factor variables in the “exposome” data frame from text to numbers.
2. Run MUVR models with all exposome data as X variables and phenotype variables (one by one) as Y. MUVR models with Q2 >0.2 were selected.
3. Selected exposures associated with at least one of the outcomes from MUVR modelling were validated using spearman partial correlations adjusted for age, sex and cohort with multiple testing FDR adjustment.

Scripts:

1. script#1\_prepping\_data.R
2. script#2\_exposure\_vs\_phenotypeMUVR.R
3. script #3\_partcor\_EM.R

Necessary data files when running scripts:

Script#1:

* “exposome.Rdata”

Script#2:

* “exposomeEdited.Rdata”

Script#3:

* “exposomeEdited.Rdata”
* “exposure\_VS\_phenotype\_models.Rdata”

Output:

Script#1:

* “exposomeEdited.Rdata” - A save file containing the edited ‘exposome’ data frame

Script#2:

* “exposure\_VS\_phenotype\_models.Rdata” - A save file containing all phenotype vs exposure MUVR models

Script#3:

* “selected\_exposures\_partcor.Rdata” - A save file containing all results from the partial correlation of exposures from VIPlists of MUVR models with Q2 > 0.2 and the corresponding phenotypes. Confounders used for the partial correlation were: Cohort, age and sex
* Two .csv files containing the exposures which had significant partial correlations for the two respective phenotypes.

1. ***Exposome-Omic Modules***

Steps:

1. MUVR models are computationally heavy and using all parameters from the methylation and gene expression data would take too long for this challenge. Therefore, sPLS modelling (the ‘spls’ function from the ‘mixOmics’ package) was used as a preprocessing step to reduce the number of variables. sPLS models were made for each exposure selected from MUVR models and which passed the partial correlation vs the methylation and gene expression data frames respectively (Methylation data was modelled with M-values). This approach reduced the two large datasets to ~1000 variables each for every exposure of interest.
2. The datasets metabol\_urine, metabol\_serum, proteome, and sPLS-filtered gene expression and methylation were combined based on sample ID and merged data sets for each exposome variable saved as a csv file.
3. MUVR models were then made for each exposure (as Y) and their respective multiOmics datasets (as X). All models with Q2 > 0.2 were then stored.

Scripts:

1. script#4genexpr\_spls.R
2. script#4methyl\_spls\_M.R
3. script#5prepare\_dataset\_multiomics.R
4. script#6exposure\_vs\_multiomicsMUVR.R

Necessary data files when running scripts:

Script#4genexpr\_spls:

* “genexpr.Rdata”
* “selected\_exposures\_partcor.RData”

Script#4methyl\_spls:

* “methy.Rdata”
* “selected\_exposures\_partcor.Rdata”

Script#5:

* “methy.Rdata”
* “metabol\_serum.Rdata”
* “metabol\_urine.Rdata”
* “proteome.Rdata”
* “genexpr.Rdata”
* “selected\_exposures\_partcor.Rdata”
* “gene\_whichElected.rda”
* “methy\_whichElected.rda”

Script#6:

* “selected\_exposures\_partcor.Rdata”
* “incommon\_id.rda”
* “multiOmicsXvars.rda”

Output:

Script#4genexpr\_spls:

* “gene\_spsModelList.rda” - A list containing all sPLS models for the exposures of interest
* "gene\_spsLoadingsList.rda" - A list of all loading of the sPLS models for the exposures of interest
* "gene\_whichElected.rda" - A list of vectors containg the sPLS loading names (genes) from the sPLS models for exposures of interest

Script#4methyl\_spls:

* "methy\_spsModelList\_M.rda" - A list containing all sPLS models for the exposures of interest
* "methy\_spsLoadingsList\_M.rda" - A list of all loading of the sPLS models for the exposures of interest
* "methy\_whichElected\_M.rda" - A list of vectors containg the sPLS loading names (methylated sites) from the sPLS models for exposures of interest

Script#5:

* “incommon\_id.rda” - A vector containing the IDs that all omics-datasets have in common
* “multiOmicsXVars.rda” - A list of multiOmics data frames corresponding to all exposures of interest

Script#6:

* “allMultiOmicsMUVRModels.Rdata” - A save file containing exposure vs multiOmics models for all exposures of interest, a list of MUVR models for all models with a Q2 > 0.2 as well as a list of the multiOmics data frames for those models.

1. ***Exposome-Omic-Outcome Modules***

Steps:

1. Spearman partial correlation to validate the selected omics in the MUVR models adjusting for age, sex and cohort and multiple testing FDR.
2. Linear modelling of the validated selected omics (from MUVR models and adjusted spearman partial correlation for age, sex and cohort) with the BMI and RAVEN outcomes.
3. Visualization in a triplot. Selected omics are aggregated in principal component analysis and individual loadings are plotted. Correlations of principal components with exposures are superimposed. Associations of principal components with outcomes are superimposed. Both correlations and associations are adjusted for age, sex and cohort. In addition, we tried a full adjusted model also adjusting for education of the mother, age of the mother, BMI of the mother, parity and year of birth. Furthermore, we tried a model additionally adjusting for two potential mediators of gestational age and weight gain during pregnancy.

Scripts:

1. 7#partcorr\_omicVars.R
2. 8#lm\_omicVars.R
3. 9#generate\_triplots.R

Necessary data files when running scripts:

Script#7:

* “finalOmicsVars.Rdata”
* “exposomeEdited.Rdata”

Script#8:

* “PartCorrOmicsVars.RData”
* “exposomeEdited.Rdata”

Script#9:

* “finalOmicsVars.Rdata”
* “exposomeEdited.Rdata”
* “metabol\_serum.Rdata”
* “metabol\_urine.Rdata”

Output:

Script#7:

* “PartCorrOmicsVars.Rdata” - A save file containing all omics variables selected from MUVR models (in script#6) which had Q2 > 0.2 and which were significant in partial correlation of exposures vs multiOmics variables.
* A .csv file for each exposure containing the multiOmics variables which were significant in partial correlation of exposures vs multiOmics variables.

Script#8:

* “finalOmicsVars.Rdata” - A save file containing two lists (one for each phenotype of interest) of multiOmics variables (from script#7) which were significant in glm analysis of multiOmics variables vs phenotype.

Script#9:

* Generates a pdf file with a triplot to visualize the results.