Metabotyping in the ”Healthy Nordic Diet” synthetic dataset

Associate Professor Carl Brunius | [carl.brunius@chalmers.se](mailto:carl.brunius@chalmers.se)

In this computer lab, you will perform a metabotyping exercise using synthetic data generated to resemble an authentic dataset reported in [1]. The synthetic data set was originally generated for a paper describing the development of a tool for molecular epidemiology and is taken from the corresponding “triplot” package [2]. In brief, this dataset was simulated from data used in a case-controlled study nested within the Swedish prospective Västerbotten Intervention Programme (VIP) cohort. The original study material was used to investigate how the plasma metabolome and the risk of developing T2D were related to compliance to a Healthy Nordic Diet. Each case was individually matched to one nondiabetic participant on age, gender, sampling date, and sample storage time. Untargeted liquid chromatography quadrupole time-of-flight mass spectrometry (LC-qTOF-MS) metabolomics was performed on plasma samples using reverse phase and hydrophilic interaction chromatography in both positive and negative electrospray ionization modes. In total, 31 plasma metabolites related to a priori-defined healthy Nordic dietary indices, i.e., the Baltic Sea Diet Score (BSDS) and Healthy Nordic Food Index (HNFI), were selected using a random forest algorithm incorporated into a repeated double cross-validation framework with unbiased variable selection (MUVR) [3,4]. Subsequently, associations were investigated between the 31 dietary index-related metabolites, dietary intakes, and T2D risk.

We assume a basic understanding of how to use R and RStudio and that you have installed relevant packages as per the instructions given prior to the NuGO course. In the lab, you will perform a series of tasks:

* Ensure installation of relevant packages
* Load data
* Make PCA and biplots
* Perform metabotyping by k-means clustering
* Associate metabotypes to baseline characteristics
* Associate metabotypes to metabolite profiles
* Associate metabotypes to diet

Following the R script “metabotyping HND.R”, using the data “HealthyNordicDiet.rda”, both on the course repository, you should be able to generate metabotypes and investigate their associations to metabolites, baseline characteristics and diet.

Please ensure that you follow the code carefully to ensure that you understand what is happening during the different stages of the script. You are strongly encouraged to think critically about the exercise!

Collect critical thoughts and questions – Start from the below and add to it:

* Is the metabotyping robust? And how can you tell?
* Are there other clustering approaches that you could have tried?   
  What happens if you use other methods?
* Is the clustering sensitive to the selection of clustering variables?
* Is the clustering sensitive to variable scaling?
* Is the clustering sensitive to potential parameter choices in the clustering algorithm?
* Should the clustering have been made on both cases and controls?
* Are the statistical tests correct?
* …
* …
* …

**References**

[1] Shi L, Brunius C, Johansson I, Bergdahl IA, Lindahl B, Hanhineva K, et al. Plasma metabolites associated with healthy Nordic dietary indexes and risk of type 2 diabetes—a nested case-control study in a Swedish population. Am J Clin Nutr 2018. https://doi.org/10.1093/ajcn/nqy145.

[2] Schillemans T, Shi L, Liu X, Åkesson A, Landberg R, Brunius C. Visualization and interpretation of multivariate associations with disease risk markers and disease risk—The triplot. Metabolites 2019;9. https://doi.org/10.3390/metabo9070133.

[3] Shi L, Westerhuis JAJA, Rosén J, Landberg R, Brunius C. Variable selection and validation in multivariate modelling. Bioinformatics 2019;35:972–80. https://doi.org/10.1093/bioinformatics/bty710.

[4] Yan Y, Schillemans T, Skantze V, Brunius C. Adjusting for covariates and assessing modeling fitness in machine learning using MUVR2. Bioinforma Adv 2024;Accepted.