Dealing with Batch Effects in Metabolomics Data: A Comparison of ComBat, WaveICA2, and a Novel Residuals Method for Classification

Blanca Rius-Sansalvador[1], *Elies Ramon*[2], *Mireia Obón-Santacana*[3], *Victor Moreno*[4]

[1]brius@idibell.cat. Unit of Biomarkers and Susceptibility (UBS), Oncology Data Analytics Program (ODAP), Catalan Institute of Oncology (ICO)[a]. ONCOBELL Program, Bellvitge Biomedical Research Institute (IDIBELL)[b]

[2]egurrea@idibell.onmicrosoft.com [a,b]

[3]mireiaobon@iconcologia.net [a,b] Centro de Investigación Biomédica en Red de Epidemiologia y Salud Pública (CIBERESP)[c]

[4]v.moreno@iconcologia.net. [a,b,c] Department of Clinical Sciences, Faculty of Medicine and Health Sciences and Universitat de Barcelona Institute of Complex Systems (UBICS), University of Barcelona[d]

Background: Metabolomics analyses small molecules in biological samples, identifying biomarkers for diseases and effects of lifestyle factors. The identification of metabolomic differences between patients of different phenotypes is key to defining biomarkers for diagnosis. Unfortunately, large metabolomics studies are challenged by batch effects resulting from diverse sources. Non-biological systematic biases need identification and correction before biomarker discovery.

Purpose: Several methodologies have been proposed to correct batch effects of metabolomics data. We aim to identify the best one, that removes unwanted systematic differences between batches but preserves biological variability.

Methods: A study on plasma metabolomics data from 510 subjects (156 with colorectal cancer or high-risk adenoma, and 354 controls). The original data was transformed with arc-hyperbolic-sine to reduce asymmetry allowing for zero values. We propose a method based on the residuals of a linear model whose covariables are the Principal Components (PCs) where is it possible to identify clusters related to the batch. The residuals of this model are scaled with the original mean and standard deviation and used as new data. This method was compared to other published ones: ComBat and WaveICA2. Visual inspection of PC plots and temporal sequence plots were used to check the effect of the batch correction. The performance of a Random Forest classification model was used to assess variability relevant to the prediction of the cancer status.

Results: The representation of original data in a PC plot showed a striking batch effect. ComBat was a slightly worse method for removing batch effect in these data, since some grouping persisted. With WaveICA2 and the residuals method, samples did not cluster in the PC plots. The performance of Random Forest classification models, after an exhaustive hyperparameter optimization, showed similar cross-validated predictive AUC (around 0.70) for the method based on residuals and WaveICA2.

Conclusions: A simple method based on the residuals of a linear model adjusted for principal components was comparable to state-of-the-art WaveICA2 to reduce batch correction and maintain classification AUC.

Keywords: metabolomics, batch effect, residuals