Strategy to remove batch effect between different omic data types

Manuel Ugidos[1,*], Sonia Tarazona[1,2], J.M. Prats-Montalbán[2], Alberto Ferrer[2], Ana Conesa[1,3]

[1] Genomics of Gene Expression Laboratory, Centro de Investigación Príncipe Felipe, Valencia, Spain
[2] Department of Applied Statistics, Operations Research and Quality, Universitat Politècnica de València, Spain
[3] Microbiology and Cell Science Dep, Institute for Food and Agricultural Sciences, University of Florida, Gainesville, USA

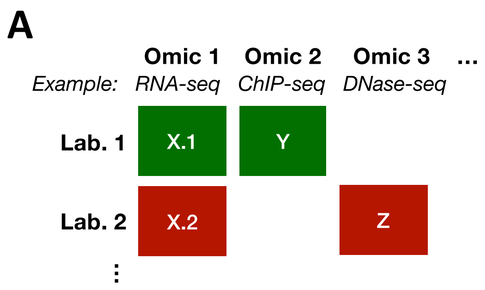
* Student information: **PhD Programme**: Statistics and Optimization; **UPV e-mail**: mauggue@doctor.upv.es

Introduction

Omics technologies have expanded in diversity in the last years and the number of omics integration analysis possibilities has also increased. However, the costs of the different techniques are still high and most of research groups cannot afford research projects where many different omics techniques are analyzed. Nevertheless, as most research share their data in public repositories, there is a possibility of utilization of datasets from other laboratories to construct a multiomic study. An important issue when we want to integrate data from different studies is the batch effect. There are already several methods described which are able to correct batch effect on common omic data between different studies (e.g. ARSyN from M.J. Nueda et al., 2012) but they cannot be used to correct no common data (i.e. the omic data modality that has been analyzed at only one lab). We have developed **MultiBaC**, a strategy to correct batch effect on no common omic information which let us integrate different omic data types from different studies.

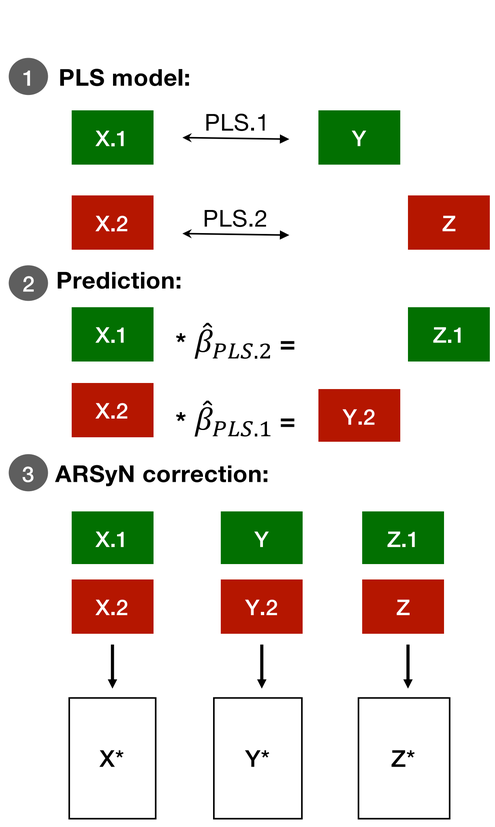
Materials & Methods

MultiBaC (**Multi**-omic **B**atch **C**orrection) method



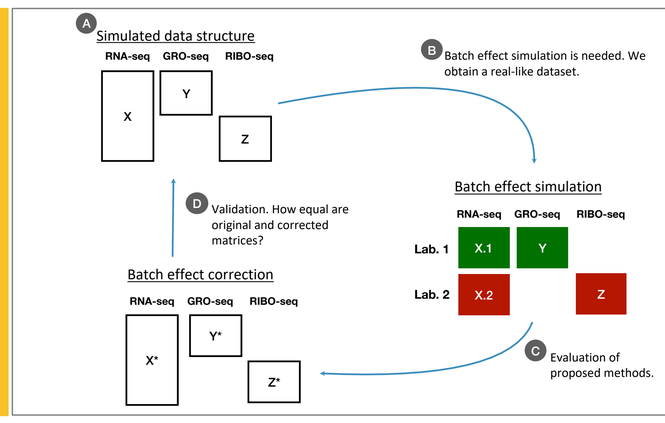
A) Minimum size problem example. At least one omic data type is shared between two laboratories.

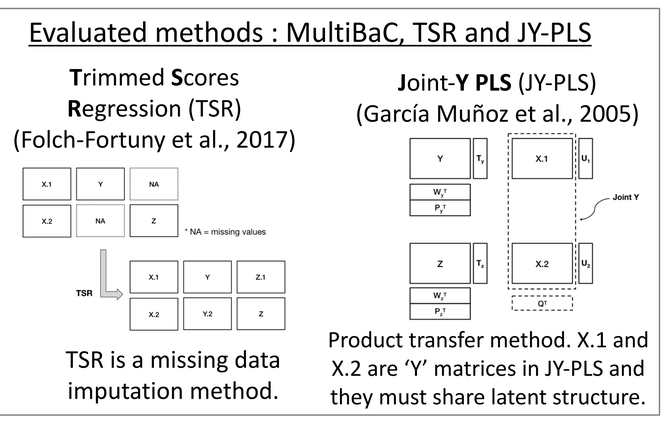
B) Method scheme. 1- A PLS model for each laboratory is built explaining non-common omic by the common one. 2- For each laboratory the omic they have not originally is predicted. 3- ARSyN correction is now possible for each omic data type.



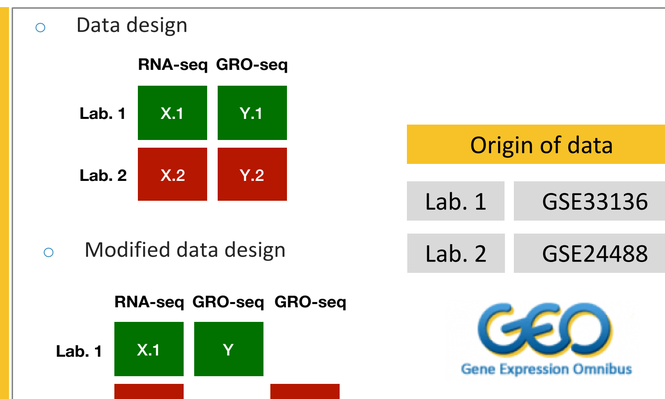
Test and Validation

Test with simulated data





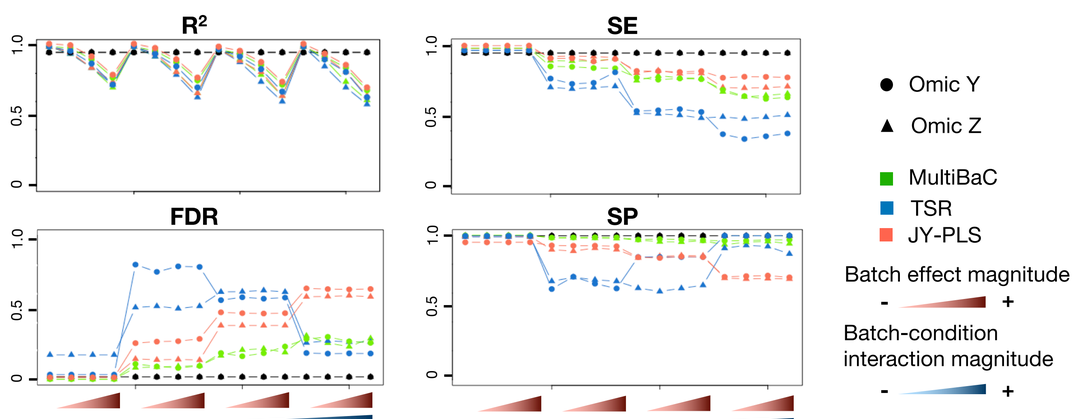
Validation with real data



- GRO-seq is being treated as no common data.
- MultiBaC is compared with ARSyN correction which is the standard method when omics are present in both labs.
- ARSyN result is the maximum objective for MultiBaC approach.

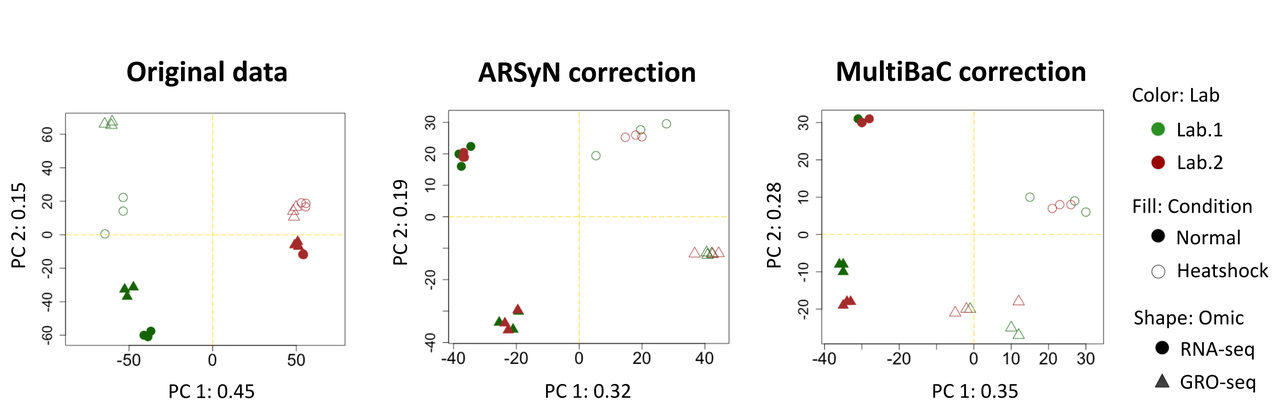
Results

With simulated data



- Measures of similarity between original simulation (no batch effect) and corrected matrices:
  - · R^2: Latent structure.
  - · FDR, Sensitivity and Specificity: Differences in significant differentially expressed genes.
- MultiBaC reaches the best performance in comparison with TSR and JY-PLS.
- Real batch effects are not supposed to be as high as maximum magnitudes tested, so MultiBaC is suitable for real cases of batch effect.

With real data



- PCAs showing sources of variability.
- MultiBaC correction is not as perfect as ARSyN one but after MultiBaC correction the batch is not an important source of variability. ARSyN result is the maximum level of correction and MultiBaC reaches almost the same result.

Conclusions

1. We have developed MultiBaC, a strategy to remove batch effect between different omic data types coming from different studies.
2. MultiBaC approach reaches almost the same performance as ARSyN method which is the maximum level of correction, when both can be applied.
3. When there is a no common omic MultiBaC is able to correct low and moderate batch effect magnitudes.
4. MultiBaC does not work well with high batch effect and interaction magnitudes but these so high magnitudes have not been seen in real datasets.

Acknowledgments

This work is part of a research project that is totally funded by Generalitat Valenciana through PROMETEO grants programme for excellence research groups.

References

[1] Maria j. Nueda, Alberto Ferrer, Ana Conesa; ARSyN: a method for the identification and removal of systematic noise in multifactorial time course microarray experiments, *Biostatistics*, 13 (2012) 553–566.

[2] A.Folch-Fortuny, F.Arteaga, A.Ferrer, Missing data imputation toolbox for MATLAB. Chemometrics and Intelligent Laboratory Systems. 154 (2016) 93-100.

[3] S. Garcia-Munoz, J. F. MacGregor, T. Kourti, Product transfer between sites using Joint-Y PLS. Chemometrics and Intelligent Laboratory Systems. 79(2005) 101-114.