AMR Abstract

reComBat: Batch effect removal in large-scale multi-source gene-expression data integration

Michael F. Adamer [1,2,], Sarah C. Brüningk [1,2,*], Alejandro Tejada-Arranz [3], Fabienne Estermann [3], Marek Basler [3], and Karsten Borgwardt [1, 2]*

[1] Machine Learning & Computational Biology Lab, Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland
[2] Swiss Institute for Bioinformatics (SIB), Lausanne, Switzerland
[3] Biozentrum, University of Basel, Basel, Switzerland
* these authors contributed equally

With the steadily increasing abundance of omics data produced all over the world, sometimes decades apart and under vastly different experimental conditions residing in public databases, a crucial step in many data-driven bioinformatics applications is that of data integration. The challenge of batch effect removal for entire databases lies in the large number of batches and biological variation which can result in design matrix singularity. This problem can currently not be solved satisfactorily by any common batch correction algorithm.
In this study, we present *reComBat*, a regularized version of the empirical Bayes method to overcome this limitation. We demonstrate our approach for the harmonization of public gene expression data (both microarray and bulkRNAsq) of the human opportunistic pathogen *Pseudomonas aeruginosa* and study several metrics to empirically demonstrate that batch effects are successfully mitigated while biologically meaningful gene expression variation is retained. *reComBat* fills the gap in batch correction approaches applicable to large-scale, public omics databases and opens up new avenues for data-driven analysis of complex biological processes beyond the scope of a single study.