Supplementary Material

MultiBaC: An R package to remove batch effects in multi-omic experiments

Manuel Ugidos, María José Nueda, José M. Prats-Montalbán,
Alberto Ferrer, Ana Conesa, Sonia Tarazona

Table of Contents
1 Overview of methods included in MultiBaC package ..................................................2
2 MultiBaC use case .............................................................................................................3
3 Comparison of ARSyNbac to other BECAs .................................................................4
  3.1 Qualitative comparison ...............................................................................................4
  3.2 Performance comparison ............................................................................................5
4 A note on MultiBaC scalability .......................................................................................6
Supplementary Figure 1. Overview of methods in MultiBaC R package. (a) ARSyNbac overview for three different options: 1) Batch effect correction, 2) Noise reduction for unknown batches, and 3) Correction when both types of unwanted effects are present in the data. In all cases, an initial ANOVA-like decomposition is performed and followed by PCA for the estimation of unwanted effects. (b) Overview of MultiBaC strategy, which combines PLS regression with conventional ARSyN batch effect correction.
2 MultiBaC use case

Supplementary Figure 2. Customized PCA score plots for complete yeast multi-omic example. Left plots show original data and right plots show MultiBaC corrected data. First and second principal components in the upper plots, and first and third principal components in the bottom plots.

The first and second principal components in the PCA score plot with the original data show a clear separation among the three batches (studies), while PCA score plots after correction show that the main sources of variability in the corrected data are the omic modality and the experimental condition.
3 Comparison of ARSyNbac to other BECAs

3.1 Qualitative comparison

Supplementary Table 1. Qualitative comparison of methods in MultiBaC R package to other BECAs.

| R package | Method    | Multi-omics | Known batch | Unknown batch | Input requirements         |
|-----------|-----------|-------------|-------------|---------------|---------------------------|
| MultiBac  | ARSyNbac  | No          | Yes         | Yes           | None**                    |
| MultiBac  |           | Yes         | Yes         | No            | None                      |
| sva (REF) | ComBat    | No          | Yes         | No            | None                      |
| sva       |           | No          | No          | Yes*          | Covariates                |
| limma (REF)|           | No          | Yes         | No            | Gaussian data             |
| RUV (REF) |           | No          | No          | Yes           | Covariates or control genes or samples. Count data. |

* The sva function creates a new variable when estimating the unwanted variation which can be used as a covariate in differential expression models. However, it does not provide a ready-to-go corrected dataset to be used for other different purposes.

** Covariates information needs to be provided to MultiBaC package but it is only used for visualization purposes and not given to the correction models.

MultiBaC method is the only BECA we could find that deals with batch effect correction on multi-omic datasets, providing there is one omic modality common to all the batches.

Regarding BECAS for the single omic situation and the known batches scenario, some of the most popular choices for batch effect correction are ComBat, with parametric and non-parametric options, and the limma package, which is based on linear models and hence requiring Gaussian-distributed data. Our ARSyNbac algorithm is based on principal component analysis and can be applied to any data type.

The removal of unwanted effects from unknown sources has been addressed by the sva method and the many variants of the RUV strategy, which are available in different R packages (ruv in CRAN, RUVnormalize and RUVseq in Bioconductor), besides the ARSyNbac methodology. In general, RUV methods need control genes or control samples to estimate the unwanted effects, which are not always available. The RUVr() function in the RUVseq package is the only one that can estimate such effects from count data without providing control genes or samples and returns a corrected dataset, as ARSyNbac does, to be used in any analysis. However, count values are required by RUVr(). The sva approach compute the surrogate variables related to unwanted noise
on any data type. These surrogate variables can be included in differential expression models but a corrected dataset to be used for other purposes is not provided.

Finally, the only available software we could verify that simultaneously removes both known and unknown unwanted sources of variation is our ARSyNbac which, also returns a corrected dataset ready to be analysed with any methodology.

It is worth to mention here the effort done my mixOmics’ team to compare these methodologies (Wang and LêCao, Briefings in Bioinformatics, 2020) and to provide R tools based on PCA to assess the relevance of batch effects. Although they have focused these studies on microbiome data, most of their findings can be generalized to other omic modalities.

3.2 Performance comparison

Supplementary Figure 3. PCA score plots for the performance comparison of different BECAs on expression data from the yeast example dataset. The two first principal components are represented in all plots. (a) Original expression data before any correction. (b) ARSyNbac corrected data with batchEstimation option (known batches). (c) ARSyNbac corrected data with both batchEstimation and filterNoise (unwanted variation from unknown sources) options. (d) limma corrected data. (e) ComBat corrected data, no covariates provided. (f) RUVEq corrected data with RUVR option. Normalized expression data was previously transformed into count data. Generalized Linear Models for residuals estimation were obtained with edgeR package following guidelines in RUVEq vignette.
The batch effect can be clearly observed for the two principal components (PC) in the PCA score plot for the original data (Fig. 3a). Fig. 3b shows that ARSyNbac batch effect correction was successful since the first PC is associated now to the experimental condition and describes around 51% of the total variability of the data. Limma correction (Fig. 3d) showed an identical performance to ARSyNbac, while ComBat (Fig. 3e) provides a similar degree of correction but, in this case, the first PC separating the experimental conditions accounts for only 48.3% of the total variability, indicating poorer performance in comparison to ARSyNbac. The best correction was obtained with ARSyNbac when using both the known batch effect option and the noise reduction (Fig. 3c), as the first PC explains 63.3% of the variability. Finally, RUVseq correction was the worst, since the first PC does not completely separate the conditions.

4 A note on MultiBaC scalability

MultiBaC method mainly relies on PCA and PLS models, which are dimension reductions methods that can handle matrices of high dimensionality, such as those generated from single-cell technologies. Moreover, both algorithms have been implemented with the NIPALS iterative procedure to avoid problems with R memory when multiplying large matrices. However, MultiBaC can be still computationally intensive for high dimensional datasets due to the cross-validation approaches needed to estimate components or predictive capacity ($Q^2$). We estimated around 90 minutes of computation time when being applied to a simulated dataset with 1000 observations (samples) per omic and batch. Supplementary Figure 4 shows the PCA score plots for this dataset before and after MultiBaC correction to illustrate the good performance of MultiBaC correction despite the dimensionality of the data.
Supplementary Figure 4. PCA score plots for a large simulated dataset (1000 samples per omic and batch). Before MultiBaC correction (original data), first principal component (PC) clearly separates batches A and B. After MultiBaC correction (corrected data), the two first PCs separate the other omic modalities, while the third component separates experimental conditions C1 and C2.