Two methods to detect batch effect applied to the comparison of batch effect correction algorithms to build Expression Atlases

Guillaume Heger

11/06/2019

I introduce here two different methods to be used in the context of the integration of datasets from different sources, subject to batch effect.

The final aim of this work is the creation of several Expression Atlases by merging the RNASeq datasets available on the existing Expression Atlas. The building of such atlases shall be done using heterogenous datasets which can definitely not be supposed to follow the same distribution. Therefore many batch effect correction algorithms can already be considered as irrelevant for our problem (batch-mean centering to name but one, as well as every method which doesn't take biological factors into account).

A sample-based detection method

A new gene-based detection method

I introduce here a new method to evaluate the importance of batch effect between datasets. This method is based on some geometric consideration on the principal components of the single datasets compared to those of their merger (with or without correction). As principal components represents geometrical directions, the best way to compare them is to estimate the angle between them.

In the simple case where datasets are supposed to have the same distribution, we expect actually the single datasets to have their principal components similar between them and to those of the merged dataset.

As the coefficients of principal components are akin to weights on the genes and represent somehow their involvement in the variance of a dataset, they are a good summary of the information provided by a dataset on the genes. Thus having low angles between the respective principal components of each dataset and the ones of their merger means that the information provided by the single datasets has been conserved through their integration.