Classnote 3:
Batch effects, technical variables, and unwanted variation

Maciej Bielecki

In molecular biology, **batch effects** refer to changes in experimental data induced by seemingly irrelevant non-biological factors, such as variance in personnel, conditions within the laboratory, or minute differences between the used reagents or instruments.

As batch effects lead to unwanted variation in the data, they can influence what conclusions can be derived from it, and thus need to be dealt with appropriately. Further confounding the matter is the fact that the multiple **technical variables** behind batch effects are typically not recorded in a verbose manner, instead being conflated into a single **surrogate variable**, usually just the processing date of the sample.

Broadly speaking, there are three stages in the statistical analysis of batch effects: exploratory analysis, downstream analysis, and diagnostic analysis [1].

During **exploratory analysis**, samples are clustered and labeled by both biological variables, which are the actual subject of the research, and technical variables (often surrogates). Based on the PCA of the data, principal components which correlate to the technical variables are identified.

During **downstream analysis**, the data is adjusted for identified batch effects, typically through clustering, linear regression or with ComBat (explained later). In the event that the recorded technical variables are insufficient in explaining all the artefacts present within the data, which is particularly likely if said variables were surrogates, surrogate variable analysis may be carried out to learn more about the artefacts.

Lastly, **diagnostic analysis** refers to re-clustering the adjusted data and confirming whether or not batch effects have been eliminated.

The aforementioned **surrogate variable analysis** (SVA) obviates the need to know in advance which technical variables cause batch effects, as it automatically estimates the sources of batch effects based on the data [2]. The most well-known method for carrying out SVA is the `sva` Bioconductor package [3].

ComBat is an algorithm for adjusting data for batch effects, designed to be robust for data with small batch sizes [4], available as part of the aforementioned `sva` Bioconductor package. This algorithm assumes an location/scale model (L/S model) for the data, defined as:

$$Y_{ijg} = \alpha_g + X\beta_g + \gamma_{ig} + \delta_{ig}\epsilon_{ijg},$$

where Y_{ijg} is the data matrix, α_g is the overall gene expression, $X\beta_g$ is a matrix of sample conditions, γ_{ig} and δ_{ig} are the additive and multiplicative batch effects respectively, with δ_{ig} further modified by noise ϵ_{ijg}. i, j and g represent specific batches, samples and genes, respectively.

The algorithm first standardizes the data, such that genes have similar overall means and variances. The standardized data is calculated as:

$$Z_{ijg} = \frac{Y_{ijg} - \hat{\alpha}_g - X\hat{\beta}_g}{\hat{\sigma}_g},$$

where Z_{ijg} is the data after standardization, $\hat{\alpha}_g$ and $X\hat{\beta}_g$ are estimates of the corresponding model variables, and $\hat{\sigma}_g$ is the estimated standard deviation.

Upon standardization, the batch effects' parameters are estimated via conditional posterior means, and the data is adjusted based on those estimates as:

$$Y^*_{ijg} = \frac{Y_{ijg} - \hat{\alpha}_g - X\hat{\beta}_g - \hat{\gamma}_{ig}}{\hat{\delta}_{ig}} + \hat{\alpha}_g + X\hat{\beta}_g,$$

where Y^*_{ijg} is data post-adjustment, while $\hat{\gamma}_{ig}$ and $\hat{\delta}_{ig}$ are estimates of γ_{ig} and δ_{ig}.

References

[1] Leek JT, Scharpf R, Bravo H *et al*, Tackling the widespread and critical impact of batch effects in high-throughput data, *Nat Rev Genet*, 11, 733–739, 2010.

[2] Leek JT, Storey JD, Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis, *PLoS Genet*, 3(9), 1724-1735, 2007.

[3] Leek JT, Johnson WE, Parker HS, Fertig EJ, Jaffe AE, Zhang Y, Storey JD, Torres LC, sva: Surrogate Variable Analysis, 2010.

[4] Johnson WE, Li C, Adjusting batch effects in microarray expression data using empirical Bayes methods, *Biostatistics*, 8(1), 118-127, 2007.