Batch effects, technical variables, and unwanted variation

Krzysztof Łukasz

High-throughput technologies have become a major tool in biological and especially genomic research. A problem related to such studies that is often not taken into account is batch effect. Measurements are affected by the reagents that are used, external conditions or staff that is performing the experiment. Unfortunately, all this can have impact on the result off the experiment.

What are batch effects?

Batch effects are variations in experiment results, that are unrelated to scientific or biological variables in a study. Different authors use different definitions, for example “the uncontrollable errors unrelated to the biological variation” , “systematic differences between the measurements of different batches of experiments” or “the cumulative errors introduced by time and place-dependent experimental variations”. They can have multiple sources. For example, they can arise due to the location of the experiment, people responsible for it, weather conditions or equipment effects.

We want to avoid batch effects, as they can lead to concerns about the validity of the conclusions drawn from the study. However, how do batch effects confound? In statistical feature selection, batch-correlated variation reduces power, with concomitant non-reproducibility. Furthermore, any classifier built using batch-correlated features does not generalize and fails subsequent independent evaluations (and as a potential diagnostic indicator). Proper removal of batch-correlated variation can remedy this, as demonstrated in both genomics and proteomics.

Batch effects correction

First step in minimizing batch effects is optimizing study design. Use of standardized protocols, maintaining laboratory conditions, use of fixed staff and reagents lot etc. can help reduce unwanted variation. In practice, batch effects are almost inevitable.

Secondly, there are computational methods that aim to minimize batch effects in samples. The classic example of a method dealing with heterogeneity of sample is normalization, a data analysis technique adjusting global properties of measurements for individual samples so they can be more accurately compared. Including a normalization step is now standard in data analysis of gene expression experiments . But normalization does not remove batch effects, because it affects different genes in different ways. In some cases, these normalization procedures may even exacerbate technical artefacts in high-throughput measurements.

Currently, there are two general classes of batch effect correction methods:

- those that use linear modeling when batches are known or assumed (e.g. ComBat)
- those that attempt to identify and control for potential batch effects (e.g. surrogate variable analysis (SVA), remove unwanted variation (RUV)).

To identify the existence of batch effects, one needs to carry out exploratory data analyses. The first step would be to perform PCA or visual methods such as hierarchical clustering. For example, if the samples cluster perfectly depending on their batch, it is highly suggestive that

differences between them are unwanted artefacts and must be accounted for in further analysis.

The most common way to deal with batch effects is to remove them via batch effect correction algorithms (BECAs). Although developed and benchmarked on microarrays in 2008, ComBat remains the most popular BECA.

Typically, the model used for genomic analysis is linear:

$$Y = XB + E,$$

where Y is a matrix of observations, X matrix of biological variables and E is the term of errors. Usually, we assume the errors to be i.i.d, but that is not necessary. We can upgrade the model to account for batch effect.

Location scale method

One of popular approaches is **Location and Scale (L/S) Adjustment**. It is a wide family of methods, assuming a mean, variance model for each batch and then adjust batches to meet the requirements. The simplest way would be to just center the means and standardize the variances across all batches for each gene. However, typically more complex solutions are needed. Thus, let us define a model where

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

α_g is overall gene expression,

$X\beta_g$ is design matrix with corresponding regression coefficients,

ε_{ijg} is normally distributed error vector with mean 0,

γ_{ig} and δ_{ig} are additive and multiplicative effects of batch *i* on gene *g* (forming L/S effect).

Now we obtain the batch-corrected data:

$$Y^*_{ijg} = \frac{Y_{ijg} - \widehat{\alpha_g} - X\widehat{\beta_g} - \widehat{\gamma_{ig}}}{\widehat{\delta_{ig}}} + \widehat{\alpha_g} + X\widehat{\beta_g},$$

where hats mean corresponding estimator based on the model.

The most popular method of the L/S family is ***ComBat***. For real application, an empirical Bayes method was applied for parameter estimation with three steps:

- standardizing the data,
- Empirical Bayes parameters estimation – additive effects assumed to be normally distributed and multiplicative coming from inverse gamma distribution,
- adjusting for batch effect.

We need to remember that ComBat assumes that the variables causing batch effect are known. Importantly, the tool can be used even when the sample is small, thanks to gathering information for a single gene from all batches. Also, the batch effect parameter estimates are shrunk towards general mean and variance (for all genes).

Surrogate Variable Analysis

SVA is an example of a tool that aims to identify unknown effects. SVA estimates the sources of batch effects directly from the high-throughput data. Variables estimated with SVA can then be incorporated into the linear model that relates the outcome to the high-dimensional feature data, in the same way as processing year or group could be included. An advantage of SVA is that surrogate variables are estimated instead of pre-specified, which means that the important potential batch variables do not have to be known in advance. Main disadvantage of this approach is that it can reduce information about subpopulation effects. Subpopulations are typically under-represented in data, and can resemble batch effect and be unintentionally removed.

After estimation, we can incorporate newly identified variables into the model, resulting in: $Y = BX + \Gamma\mathrm{G} + \mathrm{U}$, where $\Gamma\mathrm{G}$ represents the surrogate variables and U are i.i.d errors.

Estimation is based on residuals, i.e. $R = Y - \hat{B}X$. An iterative algorithm is used estimate the surrogate variables based on Singular Value Decomposition of the R matrix. It can be effectively broken down into eliminating signal form primary variables, obtaining orthogonal basis and identifying significant surrogate variables. Details can be found in (Leek & Storey 2007).

Concluding remarks

The effect that batching has had on biological research has been underappreciated. There were studies demonstrating that conclusions drawn from analyses were sometimes wrong, because the batch effect was not accounted for. It is needed to provide replicable conditions for research. There is also software to handle batch effect. When we know the variables cause=ing the variation we can use ComBat or other L/S models, when we want to estimate those variables we may use SVA, but the tools need to be used with caution, and they are far from a perfect solution.

References

Batch effect, Mikhail Dozmorov, Statistical Methods for High-throughput Genomic Data I, Fall 2017

Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007 Sep;3(9):1724-35. doi: 10.1371/journal.pgen.0030161. Epub 2007 Aug 1. PMID: 17907809; PMCID: PMC1994707.

Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010 Oct;11(10):733-9. doi: 10.1038/nrg2825. Epub 2010 Sep 14. PMID: 20838408; PMCID: PMC3880143.

Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007 Jan;8(1):118-27. doi: 10.1093/biostatistics/kxj037. Epub 2006 Apr 21. PMID: 16632515.

Goh WWB, Wang W, Wong L. Why Batch Effects Matter in Omics Data, and How to Avoid Them. Trends Biotechnol. 2017 Jun;35(6):498-507. doi: 10.1016/j.tibtech.2017.02.012. Epub 2017 Mar 25. PMID: 28351613.