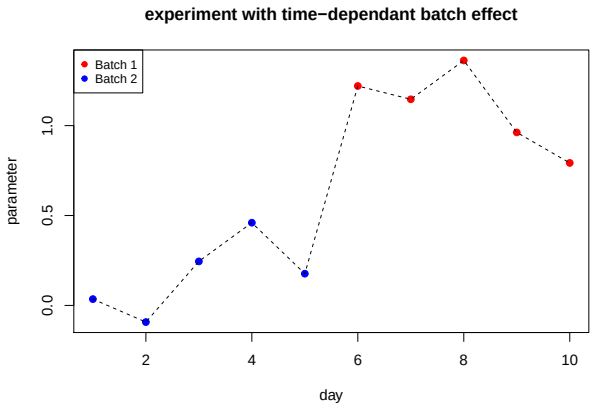
Note on Combating batch effects: surrogate variable analysis

Jan Malinowski

Batch effects emerging from differences in experiments performation. We define batch effects as a sub-groups of measurements that qualitatively vary along different conditionss and are not connected to scientific variables in a study. Such effects could emerge from differences in reagents quality or used instruments. It is popular to highlight two of such **surrogates** called:

- **processing group** – related to differences in processing protocols among laboratories and
- **date** (time) dependent effects.

Illustration of such effects was presented below using a data generated from simple $\mathcal{N}(\mu, \sigma^2)$ distribution.



Points from the first batch were generated from $\mathcal{N}(0, 0.25)$ and the latter from $\mathcal{N}(1, 0.25)$. Visible change of trend between the batches.

On this figure one can observe the change of the outcomes during the 10 days of experiment performation. We can assign the first 5 days of experiment to two "batches".

Consequences of ommiting the batch effects. In the mild cases batch effect can make the task of detecting biological signal more difficult. From the other side the strong effects could lead to misleading biological/clinical conclusions, what happened to 'home-brew' diagnostic assay for ovarian cancer, which was later blocked by US Food and Drug Administration.

What can we (analysts) do about it? Simple normalization even if obligatory to compare different features, can't overcome this type of effects. For such a task of tackling batch effects few possible statistical solutions are presented below:

- identyfing and visualizing them with PCA or hierarchical clustering techniques,
- plotting levels of individual features versus biological variables and batch variables (such as processing group or time),
- for stronger effects, downstream statistical analysis could be needed to unravel associations e.g. linear models and/or surrogate variable analysis (SVA), and finally

- good practice would also involve incorporating more "laboratory" information into analysis such as changes in personnel or reagents used in experiments.

Surrogate variable analysis (SVA). In systems biology, the phenomenon of occurence of differences in gene expression data caused by unmodeled and unmeasured factors is named **expression heterogeneity (EH)**. Potential sources of expression variation extends from environmental to demographics and genetic factors. As mentioned on previous page, acknowledgement of "batches" and their proper treatment is crucial for formulating valid conclusions from data. One method to deal with it is to use **SVA** method.

Algorithm. The method can be conceptualized in 4 steps:

1. remove the signal from primary variable(s) of interest from the data and decompose the received matrix; check if the singular vectors represent more variation than expected by chance,

2. find which genes drive each orthogonal signature of EH,

3. for each subset of genes build a surrogate variable, and finally

4. include significant surrogate variables as covariates in subsequent regression analysis to find gene-specific coefficients related to surrogate variables.

Such procedure is necessary for ensuring that *inter alia* we estimate EH and not the signal from primary variable and we take into account the fact that a surrogate variables can have different effects on individual genes.

How it compares with other methods? SVA was developed to overcome issues with properly tagging the signals from sources other than that of interest, which was not addressed by earlier analysis methods. Key difference between them and SVA is the "supervision" of factor realised by the choice of primary variables and individual **eigengenes** subsets. Additionally this approach doesn't make any assumptions about relative strenght of signal due to each source of variation.

Creating widespread dependance in expression variation across genes requires use of some adjusting techniques due to the multiple tests performed on the data. This issue is far from being simple, which means that one shouldn't adjust the EH based only on p-values and test-statistics (check Leek 2007). Moreover those classic methods of adjustement usually don't take into account the fact that the sources of dependence may generate the overlapping of the signals with primary variables of interests.

Sources:
1. Leek, J. T. Storey, J. D. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. PLOS Genetics 3, e161 (2007).
2. Leek, J. T. et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet 11, 733739 (2010).