Review of *Tackling the widespread and critical impact of batch effects in high-throughput data*

Grace Yi Chen

The paper this week discusses the impact of batch effects in high-throughput genetic data analysis. The measurements of high-throughput genetic data are affected by both biological and non-biological factors. Batch effects refer to those non-biological factors which includes differences in machine, reagents and personnel. These effects can have a big impact on the downstream analysis which would lead to questionable biological conclusions. One way to quantify the impact of batch effects is to examine the principal components of the data. The authors examined the extent of batch effects for eight published data sets using linear model and PCA and found batch effects have a large influence on data measurements. The authors propose several solutions like careful study design and distribute potential sources of non-biological factors across biological groups. We could also use statistical solutions to control like incorporating potential batch effects in the model and identifying unmeasured batch effects using surrogate variable analysis (SVA).

This paper is written well and easy for me to understand the general idea and impact of batch effects in high-throughput genetic data analysis. Batch effects are sometimes overlooked but it has a big impact on the downstream analysis. As mentioned in the paper, batch effects can confound the association between genetic variants and a phenotype of interest. For example, if samples from different batches have different gene expression, we cannot make conclusion about the association between variants and the phenotype. The authors later mentioned several ways to control batch effects. When using SVA, sometimes it is hard to interpret surrogate variables as they may not have a meaning. Maybe it is useful to compare the results of SVA to other methods controlling batch effect for sensitivity analysis.

Question:

1. I am wondering if there are any new methods identifying and controlling batch effects?