Summary

This paper notes that batch effects in a lot of assays are a major confounder for the biological processes they are for. This is important because the batch id, the time the batch was measured, or the equipment used for a batch are assumed to not affect the outcome in most models. This paper warns of these effects and says that information about batches measured should be recorded as well.

Reaction

Seems like a very serious problem that should always be considered in biological data analysis. They do a good job of showing how much of an issue this is by showing how the data groups up on these confounders. It might have been nice to know why this happens (why does the time a batch was taken matter so much).

Questions

I didn’t quite follow their logic with PCA here:

“Neither date nor biological factors was perfectly associated with the top principal components, suggesting that other unknown sources of batch variability are present.”

Isn’t this to be expected, why would that suggest other batch sources of variability?

2.

Would normalizing for PCA also violate their independent batch assumption?