The Practical Effect of Batch on Genomic Prediction

Hilary S. Parker, Jeffrey T. Leek

Department of Biostatistics, Johns Hopkins School of Public Health

Abstract

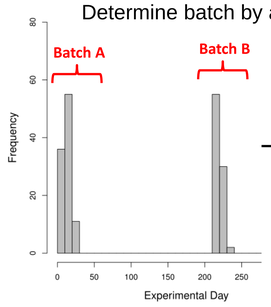
Measurements from microarray and other high-throughput technologies are susceptible to a number of non-biological variables such as temperature and reagent lots. It has been shown that these artifacts, collectively called batch effects, can severely alter the outcome of any differential expression analysis, resulting in misleading biological conclusions. Here we examine the impact of batch effects on predictors built from genomic technologies - specifically gene expression microarrays. We compare single microarray (fRMA) and multiple microarray (RMA) preprocessing methods and both rank-based (top-scoring pairs) and continuous (PAM) predictors. We show that in general, prediction is made more difficult by batch effects. We also show that when there is perfect confounding of batch and the outcome being predicted, then accuracy is substantially reduced. In an effort to mitigate this effect, we determine which probes from commonly used Affymetrix arrays are most susceptible to batch and investigate their properties. Down-weighting these "batch-affected" probes may lead to increased predictive accuracy when building gene expression based predictors.

Data Example

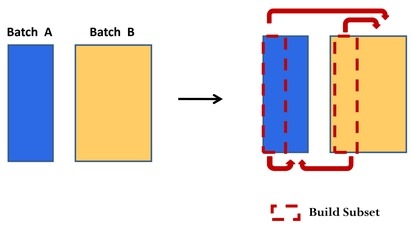
Design:

To test the effect of batch on prediction, we examined the cross-validated accuracy of prediction in large datasets with independently distributed outcomes and batches.

Determine batch by array date

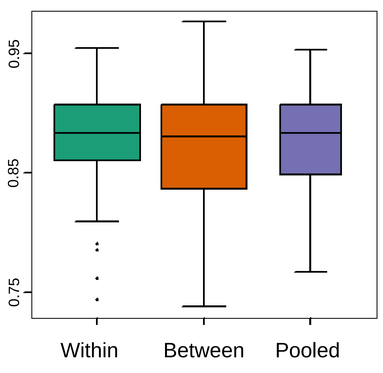


Build models on "build" subset, and test model on remaining data both within and between batches.



Results: Cross-Validated Accuracy

Models built on the same batch perform better, and with less variance, than models built on separate batches.

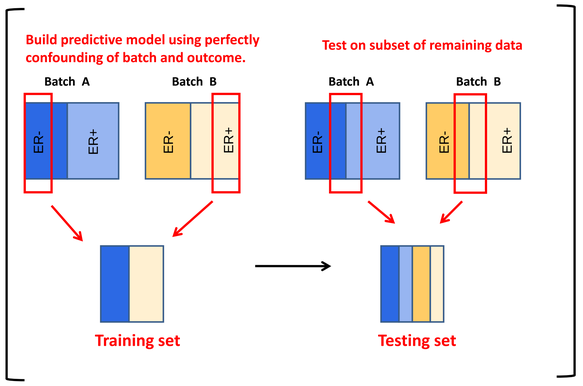


A typical result: Within-batch cross validation accuracy is slightly higher than the between-batch cross validation success rate. The variance is also higher with between-batch prediction.

Perfect Confounding

Design:

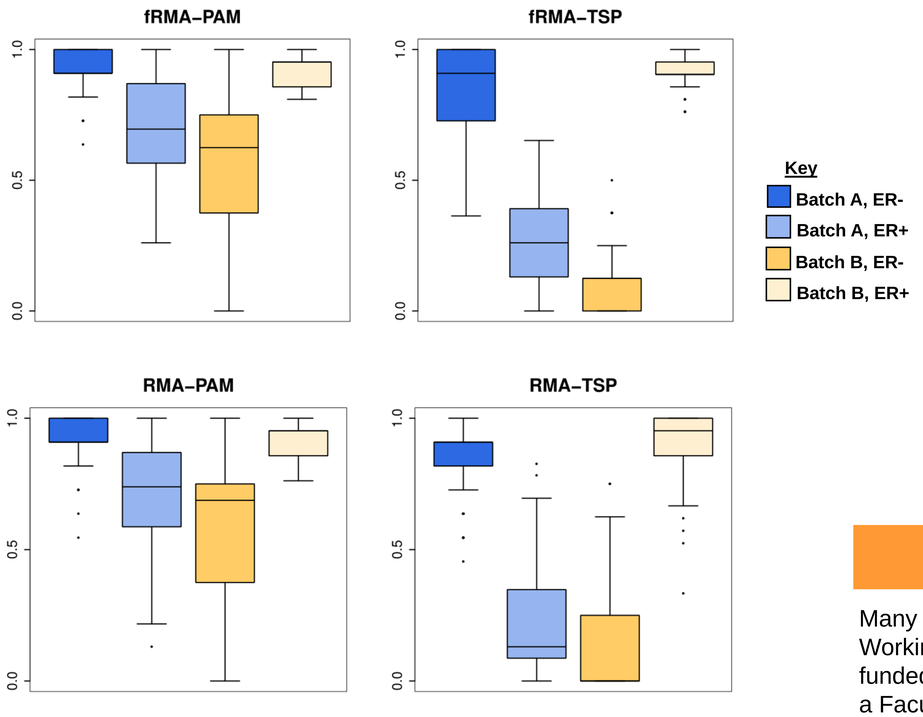
To test the effect of perfect confounding between batch and outcome on prediction, we build a prediction model on an artificially perfectly confounded dataset, and test performance on non-confounded outcomes.



× 100

Results: Cross-Validated Accuracy

A preprocessing methods designed for single arrays (fRMA) combined with a predictor using several probes (PAM) was least affected by batch. A rank-based prediction algorithm (TSP) performed similarly using data normalized with both single-array (fRMA) and multi-array (RMA) methods.



Correcting for Batch

Design:

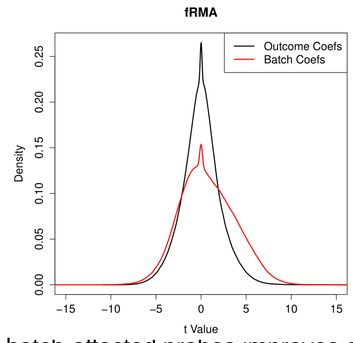
To determine whether certain probes are more batch-affected than others, and furthermore whether removing these probes improves cross-validated accuracy, we first fit the following model on the non-testing subset of the data:

$$Y_{ij} = \beta_{0i} + \beta_{1i}\text{batch}_j + \beta_{2i}\text{outcome}_j + \varepsilon_{ij} \qquad (1)$$

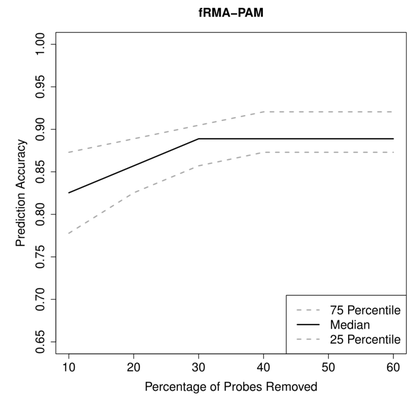
for probe set i on array j. We determined batch-affected probes by ranking the β_{1i} p-values and removing the top quantile.

Results:

Density plots for model 1 reveal differing distributions for batch and outcome parameters.



Removal of batch-affected probes improves cross-validated accuracy in all combinations of preprocessing method and prediction algorithm. fRMA-PAM combination shown.



Acknowledgements & Contact

Many thanks to Rafael Irizarry and the Genomics Working Group at Johns Hopkins. This work was funded by the JHSPH Sommer Scholar program and a Faculty Innovation Award to Jeffrey Leek.

hiparker@jhsph.edu

www.biostat.jhsph.edu/~hiparker/PracticalBatch