

Using Statistics to Improve the Quality of Genomic and Proteomic Data

Simple method reduces noise in measurement platforms

Pharmaceutical and biotechnology companies today are investing billions of dollars collecting, storing, and analyzing data from new genomic and proteomic technologies. It is expected that analysis of this data will result in the discovery of new biomarkers and improve the efficiency of the drug discovery process. If this new technology is to live up to this expectation, it is vitally important to be sure that the analyzed data is of the highest quality.

Identifying sources of variability

Although many believe the data from these new technologies to be noisy and relatively non-reproducible, research shows that, with a little extra effort, the measurements from these new technologies can indeed be very reproducible. Over the past few years, my organization has participated in more than two dozen microarray studies that attempted to identify the major sources of noise in microarray data. The results were surprising. In many cases, the effects such as sample preparation batch, reagent lot, and microarray lot were larger than the biological effects being studied. This leaves the researcher in one of two situations:

1. If the processing batches are confounded with the treatments, the treatment effects will seem larger

than they really are and will lead the researcher to incorrectly conclude that the treatment has a large effect in gene expression when in fact the effect they are seeing may be a batch effect.

2. If the processing batches are not confounded with the treatments, the variability due to the batch effects may obscure the treatment effects.

When a study does not benefit from sound experimental design, we often find ourselves in the first situation — nuisance effects are confounded with the treatments. For example, suppose the effects of a drug treatment in mice are being studied, and the treatments include a combination of treated vs. control at day 0, day 3 and day 7. If the samples from day 0 are all processed in one batch, day 3 in another batch, and day 7 in a third batch, the processing batch is confounded with the time point. This design will not allow the researcher to determine whether the differences seen are due to changes in time or whether they are due to the processing batches.

A better approach is to include the processing batches in the experiment design, allowing the researcher to distinguish between batch effects and treatment effects. This “randomized block” design, in which each batch contains

In many cases, the effects such as sample preparation batch, reagent lot, and microarray lot were larger than the biological effects being studied.

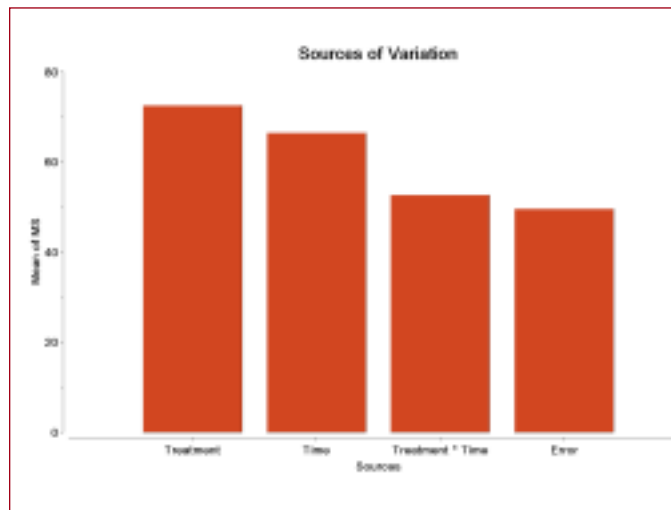


Figure 1a: Relative sizes of treatment effects averaged over ~20,000 genes. Note the substantial size of the error (noise).

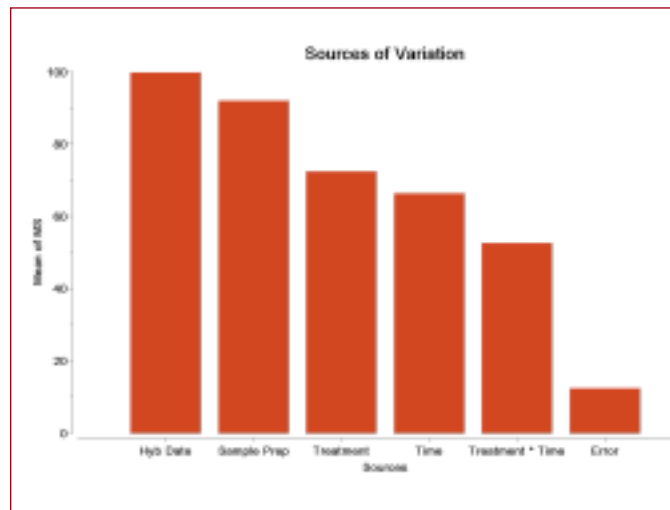


Figure 1b: Sources of variation as estimated using a four-way analysis of variance that includes sample preparation batch and hybridization date. Note the substantial reduction in noise achieved by including the batches in the analysis.

each of the treatment combinations in balanced proportion, is critical in producing high-quality data from today's new genomic and proteomic measurement technologies.

Identifying and removing batch effects

In the following example, a typical microarray experiment is used to demonstrate how batch effects can easily be measured and removed from the data, revealing the true treatment effects. In this experiment, a drug treatment is administered and samples are collected at day 0, day 3 and day 7. In addition to treatment and time, the sample preparation batch and hybridization date also were included in a randomized block design. Figure 1 shows the sources of variability in the experimental data (as determined using analysis of variance). Note that a simple analysis that ignores the processing batches leaves a lot of noise in the data (the right-most bar labeled "Error" in Figure 1a) whereas an analysis that includes the processing batches greatly reduces the noise in the data (Figure 1b).

Analysis of variance is great at partitioning sources of variability in a properly designed experiment, but what about popular multivariate methods such as cluster analysis and principal components analysis (PCA)? While batch effects tend to affect the value of all genes or proteins measured, a particular treatment or phenotype may affect only a relatively small percentage of the genome or proteome. Thus, when using multivariate techniques that look at all genes or proteins simultaneously, the batch effects often obscure the treatment effects.

Figure 2 shows the data from this same experiment

There is only one way to truly know if you have batch effects, and that is to design the experiment in such a way that these effects can be measured.

visualized using PCA. Figure 2a shows that the hybridization date is the largest effect in the data — the points largely cluster into groups of samples hybridized on the same date. Because the treatments were balanced across batches in the experiment design, the size of the batch effects can be reliably estimated and removed from the data. Figure 2b shows the data from this experiment after adjusting for batch-to-batch differences. In this figure, the samples cluster by the treatment groups, and within each treatment group, they cluster by the time point.

A simple recipe for collecting quality data

Here is a simple three-step method that you can use to improve the quality of the data from today's new genomic and proteomic measurement platforms:

1. Incorporate sample preparation and processing batches into the experiment design.
2. Estimate the batch effects using a statistical technique such as analysis of variance.
3. Adjust the data to remove the batch effects.

Many researchers ignore batch effects, believing that they are not significant. But there is only one way to truly know if you have batch effects, and that is to design the experiment in such a way that these effects can be measured. In my experience, once a researcher applies this strategy to a genomic or proteomic experiment, they will never do another experiment again without the benefit of a sound experiment design.

Tom Downey is President of Partek Inc. He may be contacted at sceditor@scimag.com.

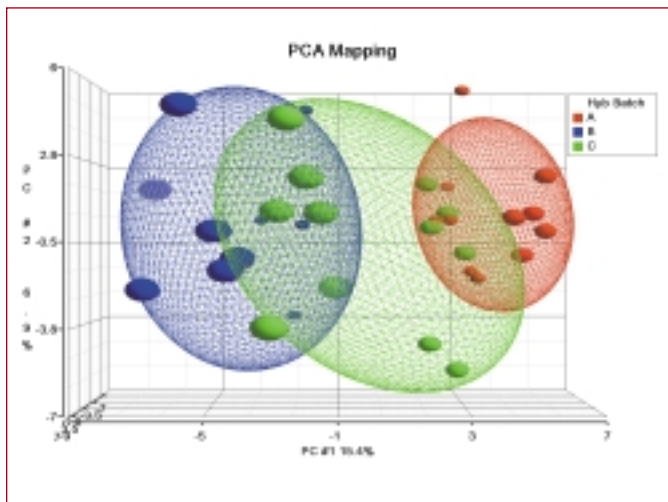


Figure 2a: PCA mapping of microarray experiment of 36 samples using Partek Pro. Samples are colored by hybridization date.

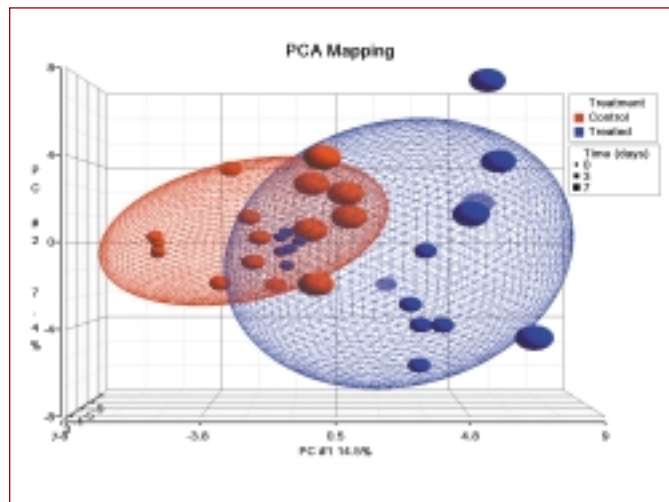


Figure 2b: PCA mapping of the same data after batch effects are removed. Samples are colored by treatment and sized by time point. Note that, within each treatment group, there is a right-to-left trend in time.