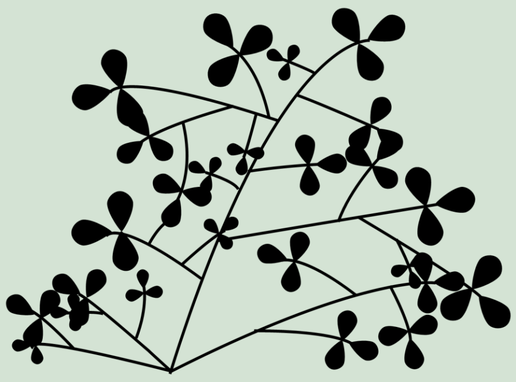
Batch Correction without QCs

Martin Rusilowicz
Julie Wilson
Simon O'Keefe
Adrian Charlton
Michael Dickinson

Background

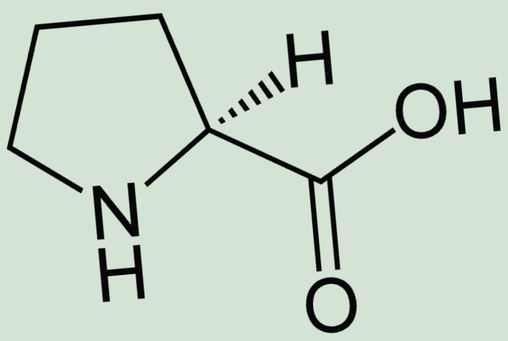


Drought and **disease** have a significant impact on crop production worldwide. Abiotic stresses can be compounded by biotic stresses, such as infection with the *Fusarium* pathogen. Here we subject *Medicago truncatula*, a model legume, to individual and combined stresses:

- Control
- Fusarium
- Drought
- Drought + Fusarium

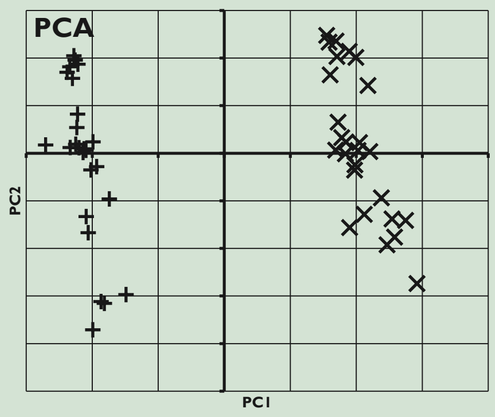
Three plants were extracted from each experimental group at daily intervals for 12 days. Leaf and root extractions were analysed using positive and negative mode LC-MS in 7 batches.

LC-MS



Unwanted **variation** is introduced into LC-MS data from a number of sources. A widely implemented solution is the inclusion of **quality control** (QC) samples into the study. These provide a fixed reference point by which any instrumental variation can be tracked.
In untargeted studies QCs typically consist of pooled experimental samples. Should insufficient material be available for pooled samples, as is the case in our study here, then biologically similar samples may also act as QC samples.

QC Correction



Changes due to **batch differences** are readily apparent in the PCA plot (*left*), which represents the peak intensities for the first two batches (shown as + and ×). Using the QC samples allows us to measure the drift in intensity of each individual peak. We trialled two methods to determine the drift of each peak for each batch:

- Mean QC Intensity
- Linear regression of QCs

Whilst these methods worked well on most of our datasets we found datasets for which these corrections made the differences **worse**. What went wrong?

Whole Data Correction

Three factors can be observed that adversely affect correction using QCs:

- Changes occurring *between* QCs
- *Intensity differences* between QC and experimental samples
- *Insufficient* QCs to track more complex baselines

Instead, we used the flow of the **whole set** of experimental data to form the correction "baseline", applying a number of smoothing functions to the data. These functions included the moving median, LOESS, splines and polynomial regression.

The figures to the left show the effects of both methods (linear-regression of QCs and moving-average of whole-data) on the same batch.
x = *QCs* ● = *Other data points* — = *Baseline used for correction*

Results

Relative Standard Deviation of biological replicates, as well as PCA plots, indicate better correction is achieved using whole-data based methods for our dataset. The simple moving average offered as good a correction as the more complex smoothing methods, likely due to its ability to rapidly track abrupt changes of instrumental drift. Applying this correction allowed batch differences to be removed as a major source of variance, emphasizing true differences between experimental samples and facilitating the identification of metabolites key to the stress response.

University of York
fera The Food and Environment Research Agency
EPSRC Engineering and Physical Sciences Research Council