Within- and cross-batch correction of a large-scale metabotyping study using a rapid FIA-MRMS profiling approach

Nikolas Kessler[1], Berin A Boughton[2], Torben Kimhofer[2], Aiko Barsch[1], Matthias Witt[1], Melvin Gay[3], Jeremy K Nicholson[2];
[1]Bruker Daltonics GmbH & Co.KG, Bremen, Germany; [2]Australian National Phenome Centre, Murdoch University, Murdoch, Australia; [3]Bruker PTY Ltd., Preston, Australia

Introduction

Metabotyping research relies on precision measurement of chemical species, with statistically well-powered studies typically comprising hundreds or thousands of samples. Robust analytical hardware, high-throughput-optimised laboratory procedures and the inclusion of effective quality control measures represent major factors for the acquisition of trustworthy data.

Data generated in large analytical runs or in batch mode often show systematic intensity drifts over time or step-function-like intensity patterns, respectively. These effects are typically removed at post-acquisition stage using statistical techniques. Here, we present a workflow that provides both, within-batch correction (WBC) addressing intensity drifts, and a custom-made cross-batch correction (CBC) that allows to append separate batches for joint analysis. The data workflow is illustrated using a flow-injection analysis (FIA) MRMS approach applied to serum samples of a SARS-CoV-2 patient cohort.

Methods

Sample preparation

A total of 589 patient sera were prepared according to standard operating procedure comprising 1:50 sample dilution and methanol extraction (95% MeOH/5% H_2O, 0.1% formic acid). Pooled quality control samples (PQC, n=77) were included for each acquisition batch, represented by a 96-well plate.

Data acquisition

A Bruker SolariX 2XR 7T MRMS calibrated with NaFormate (Δ m/z <1ppm) was used to measure samples in positive ion mode with broadband detection, averaging 32 spectra across *m/z* 100–3000 at est. resolving power 560k *m/z* 400. Samples were delivered via a Bruker Elute HT and PAL-RSI with a 20 µL sample loop installed using a 3.5 min method with an acquisition period of 1.8 min and a flow rate of 0.01 mL/min during acquisition.

Data processing

A preliminary version of Bruker MetaboScape® 2023 was used to generate a feature table, with feature being defined as isotope and adduct-deconvoluted ion signals. Features present in less than 66% of samples were excluded. Intensity drift-effects were corrected using the WBC routine implemented in the MetaboScape software.

Batch-related, systematic intensity patterns were corrected using a custom CBC routine implemented in Python at the Australian National Phenome Centre, using the MetaboScape REST-API to programmatically load feature tables from the MetaboScape processing computer (Windows 10) into the data scientist's computer (macOS).

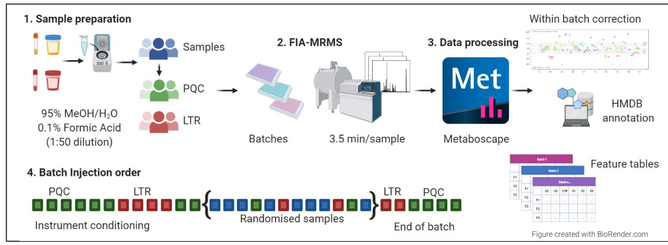

Fig. 1 – 1) Analytes are extracted by 1:50 dilution with organic solvent, cooled for 15 min and centrifuged. A PQC sample is generated from aliquots of all samples and an additional long-term reference (LTR) sample is prepared. Prepared samples may be stored at -80°C. 2) 96-well plates are run as independent batches using a FIA-MRMS approach in 3.5 mins/injection. 3) Resulting single spectral data are processed using T-ReX® 2D algorithm in MetaboScape comprising of peak picking, alignment, integration, within-batch correction and putative metabolite annotation. 4) The injection order of each batch starts with a series of burn-in PQCs/LTRs, followed by randomized sequence of patient samples with periodically inserted PQCs/LTRs, and ends with additional LTR/PQC samples. PQCs runs at the start, in between study samples and at the end are essential to calculate within batch and cross batch corrections.

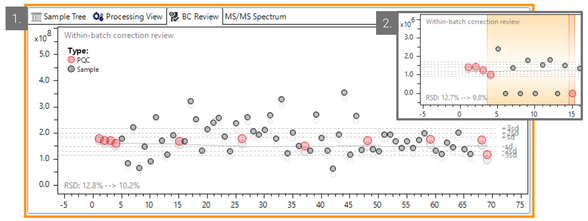

Fig. 2 – 1) Interactive review of the feature-wise within-batch correction for compound 304.2613 m/z from batch two (see Fig. 3). The intensity in that batch is very stable, however WBC slightly improves the RSD in PQCs from 12.8 % to 10.2 %. The WBC correction function is depicted as a gray line; 2) Outliers are highlighted according to the rules presented by Gika *et al.* [1], here marking one PQC outlier outside the three standard deviations (σ) range and two consecutive PQC outliers outside the two σ range.

The within-batch correction (WBC) in MetaboScape makes use of PQC samples to estimate run-order dependent intensity drifts. For every detected feature, a LOESS correction function with a large bandwidth (to avoid overfitting) is calculated, based on the PQC's intensities. This non-linear, smooth estimate of intensity drift is used to adjust feature intensities of all study samples. The WBC will only interpolate. It is thus mandatory that PQCs are run in the beginning and end of each batch.

In MetaboScape, the interactive *BC Review* plot shows both the uncorrected (in grey) and corrected intensities (color according to sample) for a selected feature. The LOESS correction function is depicted as a grey colored line (see Fig. 2.1). PQC outliers are detected according to Gika *et al.* [1] and are highlighted in orange (see Fig. 2.2). The WBC also allows to filter features based on the relative standard deviations (RSDs) in the PQCs' intensities. Here, all features that exceeded 40% RSD in the PQCs before correction or 20% RSD after correction are excluded. An additional option to discard features that were not present in all PQCs is available but here not activated.

$$RSD\ (\%) = \frac{\sigma}{\mu} * 100 \qquad \mu - \text{mean},\ \sigma - \text{standard deviation}$$

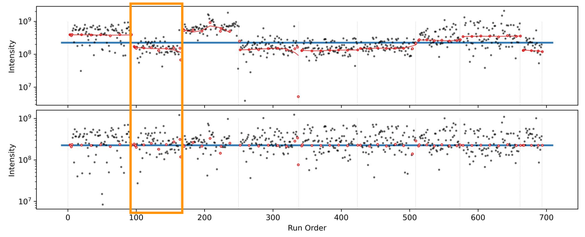

Fig. 3 – Upper panel: Raw feature intensity (m/z: 304.2613), with red-colored points indicating PQC samples and red-colored lines indicating a smooth trend estimate of drift effects for each batch (WBC).
Lower panel: Feature intensities after WBC and CBC correction. The blue line indicates the grand median of all study samples prior to any correction applied. The orange box indicates the batch displayed in Fig. 2.

Batch-related, systematic intensity patterns are removed using a custom Python routine that accounts for differences in intensity magnitudes, as well as for magnitude-dependent dispersion effects. The latter acknowledges the heteroscedastic nature of MRMS data. The mean-dispersion relationship is estimated parametrically using all available feature bucket data.

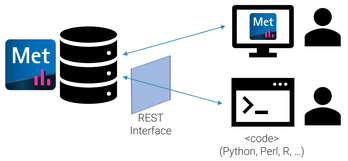

Fig. 4 – The CBC was implemented at ANPC using Python, connecting to the MetaboScape server's REST API. This interface can be easily created for any common programming language [2], including Python, Perl, R, javascript, Java, C++ and many others.

Summary

The presented workflow comprises all steps for a robust and controlled processing of biofluid samples collected in large-scale metabotyping studies: From careful design of sample plates, to data acquisition, feature extraction, and statistical removal of systematic within and cross-batch effects.

The processed feature tables from MetaboScape were accessed via the software's REST API. This new interface allows to build custom workflows that enables to combine proprietary with in-house developed algorithms. Here, this was exemplified using a combination of MetaboScape's within batch correction (WBC) and custom developed cross-batch correction (CBC). WBC and CBC are based on pooled quality control samples and ultimately allow to merge different acquisition batches for joint down-stream analyses. Statistically powerful large-scale analyses thus become feasible.

In summary, we present means to further improve data quality in large-cohort studies comprising multiple batches. MetaboScape's REST-API offers new flexibility in developing automated, instrument & assay-tailored data analysis workflows. This has been particularly useful for studies that followed non-standardized designs, such as in early stages of the COVID-19 pandemic.

Acknowledgements

We thank The Spinnaker Health Research Foundation, WA, The McCusker Charitable Foundation, WA, The Western Australian State Government, and the Commonwealth Department of Health for funding the Australian National Phenome Centre for this and related work via its Medical Research Future Fund (MRFF) Accelerated Research. We thank Paul Lyons, Ken Smith and their teams from the University of Cambridge for providing COVID-19 patient material and sample annotation data.

References

[1] Gika *et al.*, Journal of Chromatography B, Volume 1008, 2016

[2] https://editor.swagger.io (last accessed 27/05/2022)

Conclusion

- We present a workflow for the analysis of large-scale metabotyping studies, including quality control and assessment of batch effects.
- The within-batch correction in MetaboScape uses pooled quality control samples to correct for run-order dependent intensity drifts.
- A custom cross-batch correction has been implemented, to account for differences in intensity magnitudes and magnitude-dependent dispersion effects.
- The custom cross-batch correction, implemented in Python at the ANPC, directly retrieves data from MetaboScape using its REST API, which makes MetaboScape data accessible to data scientists.

MetaboScape