NPC Project Data Summary Report for: AIRWAVE

1. Abstract

This report summerises the datasets generated by the National Phenome Centre as part of the AIRWAVE project. AIRWAVE consisted of 3000 samples of serum and 3000 samples of urine.

Datasets Generated

8 assays were performed, generating 9 datasets.

Sample Type	Assay	Final Number of Samples	Number of Features Detected
Urine	RPOS	2958	14300
Urine	RNEG	2960	14481
Urine	HPOS	2982	7325
Blood Serum	LNEG	2981	5833
Blood Serum	HPOS	2975	1505
Blood Serum	LPOS	2971	7211

Urine	NMR	3000	20000
Blood Serum	NMR	3000	20000 (spectral size)
Blood Serum	BI-LISA	2999	105

2. Sample Handling

Upon receipt, all samples were unpacked from their shipping container and stored at -80°C. Samples were sorted into batches of 80 using metal racks in contact with dry ice, ensuring the samples were maintained at a temperature of -20°C or below at all times. The sorting order was determined by orthogonalisation with respect to the study design provided to the NPC. Samples were returned to -80°C until needed for sample aliquoting.

For aliquoting, samples were thawed to 4°C overnight in metal racks providing an equal rate of sample thawing, and the liquid contents were transferred to a 96-well plate. Each plate was centrifuged to remove particulate matter, and the supernatant was aspirated and dispensed across a dedicated 96-well plates for each assay. Plates were heat-sealed and returned to -80°C until needed for assay-specific sample preparation.

3. Analytical Methods and Coverage

All NPC assays include repeat measurements of a set of reference samples in addition to any assay-specific calibration samples outlined in the publications referenced below. Study-Reference (SR) samples consist of a pool of all samples for each matrix in the study, and provide a baseline representing every chemical species detectable in the study. This allows for correction of longitudinal drift in analytical measurement as well as the calculation of technical precision metrics for each feature detected. Long-Term Reference (LTR) samples are a pool of samples of an identical matrix (urine, plasma, or serum) external to the study which is used internally for quality control purposes.

Nuclear Magnetic Resonance Spectroscopy

Nuclear Magnetic Resonance Spectroscopy (NMR) based metabolic phenotyping provides a highly robust, repeatable and precise platform for the detection of small molecules in human biofluids. NMR is a strongly quantitative experimental method, in which the area of a resonance is directly related to the number of nuclei generating the signal. Therefore, the area of a resonance attributed to a known standard, either spiked into the sample, or an endogenous component quantified by other means, can be directly related to the chemical formula of that compound, and from there to the formula and signal-area of unquantified compounds. In controlled situations, components of a solution can be quantified with an error of below 1%, down to components comprising less than 0.1% of the solution.

Acquisition of NMR profiles was conducted according to the protocols laid out in Dona *et al.*. (2014). Samples are prepared and analysed daily in batches of 80 study samples with the addition of 4 quality controls (2 SR and 2 LTR). Samples were maintained at 4°C during preparation for, and while awaiting, acquisition.

Lipoprotein parameters were generated by the Bruker *B.I.-LISA* (Bruker IVDr Lipoprotein Subclass Analysis) platform. *B.I.-LISA* determines cholesterol, free cholesterol, phospholipids, triglycerides, apolipoproteins A1, A2, B and particle numbers for the primary plasma and serum lipoproteins and their subclasses.

Spectra were acquired at 600 MHz with Bruker Ascend 600 magnets and Avance III HD consoles configured to the Bruker IVDr specification (Bruker Corporation, Billerica, MA, USA).

Ultra-Performance Liquid Chromatography - Mass Spectrometry

Mass Spectrometry (MS) provides a highly sensitive platform for generating quantifications of small molecules, and when combined with the appropriate Ultra-Performance Liquid Chromatography (UPLC) provides a wide coverage of biologically relevant metabolic classes. Sample preparation for acquisition of profiling datasets by UPLC-MS was conducted in batches of 80, daily, to minimise the effect of sample aging.

UPLC-MS acquisitions were structured according to the protocols laid out in Lewis *et al.*. Samples were acquired in batches of up to 1000 study-samples, interleaved with alternating SR and LTR samples every five injections (16 per 80 samples), each batch was flanked by a serial dilution of the SR sample to assess linearity of response.

The NPCs optimised Hydrophilic interaction chromatography (HILIC, denoted H in NPC assays), provides enhanced separation of small, highly polar molecules, ionised in both positive mode, and is described in Lewis *et al.* (2016). The NPCs reversed-phase chromatography (denoted R in NPC assays) targets small-moderately polar molecules, ionised in both positive and negative modes, and is described in Lewis *et al.* (2016). The NPCs lipid-targeted reverse-phase chromatography (denoted L in NPC assays) provides maximal resolution of fatty-acids, triglycerides, and phospholipids, ionised in both positive and negative modes. Lipid-targeted chromatography is conducted according to the protocol described by Sarafian *et al.* (2014).

All UPLC-MS profiling assays were acquired on Waters G2-S ToF mass spectrometers, with Acquity UPLC chromatography systems (Waters Corporation, Milford, MA, USA).

4. Data Processing

Nuclear Magnetic Resonance Spectroscopy

NMR spectra were automatically processed in TopSpin 3.2, followed by a suite of in-house scripts, according to the protocols laid out in Dona *et al.*. Each spectrum was automatically checked to ensure analytical quality, before all spectra were aligned to a common reference scale. This matrix was then checked for multivariate outliers.

Analytical quality is assessed on 4 factors:

- Line width of less than 0.9 Hz
- Quality of water-suppression - the residual water signal did not impinge on the surrounding spectrum
- Even baseline signal - the baseline was even and flat across the entire spectrum
- Accurate chemical shift referencing. Urine samples were referenced to an internal spiked standard 3-(trimethylsilyl)-2,2,3,3-tetradeuteropropionic acid (TSP) at 0 ppm. Serum and plasma samples were referenced to the α-anomeric glucose doublet at 5.233 ppm

Quality assessment was performed daily, and any spectrum failing any of the above tests was immediately rerun. If the sample did not pass a second time, it was considered a biological outlier and was excluded from further analysis. Spectra that passed the analytical checks were aligned to a common reference scale, running from 10 to -1 ppm, and interpolated onto a common 20,000 point grid.

Lipoprotein parameters were validated according to Bruker's *B.I.-LISA* protocols.

Ultra-Performance Liquid Chromatography - Mass Spectrometry

Waters format instrument raw files were converted to .mzNLD for retention-time alignment and feature detection in Progenesis QI (Waters Corp. Milford, MA, USA). Progenesis QI was configured to align retention time to the central long-term reference sample of the acquisition. Peak detection is configured with a minimum chromatographic peak width of 0.01 minutes, and automatic noise detection set to the minimum threshold of 1. Peaks arising from isotopes and chemical adducts are automatically resolved according to the observed m/z and chromatographic peak-shape, and peaks areas integrated.

Further processing and filtering of UPLC-MS profiling datasets was conducted with a suite of in-house scripts, and used to account for analytical run-order effects and remove noise from each dataset.

Analytical run-order effects were accounted for with an adaption of the method described in Zelena E, *et. al.*. (2009). A robust LOWESS regression was generated per-feature, based on the study reference samples, in run-order, with the window scaled to include 21 SR samples. The smoothed response values for each feature were then interpolated to the intermediate study sample injections using simple linear interpolation. Finally the median intensity of each feature in each analytical batch was aligned.

Extracted features spuriously arising from analytical noise were removed from the dataset by a pair of approaches, both applied on a per-feature basis. First, a serial dilution of the study reference sample was used to asses the linearity of responses of each feature. This consisted of a linear series of dilutions of the study reference sample, run with each analytical batch. Detected features were correlated to their expected intensity in the dilution series, and those features showing an Pearson's r of less than 0.7 were excluded from further analysis. Second, the relative standard deviation (RSD) of each feature across the study reference samples was calculated, and those features where the RSD exceeded 30%, or the observed biological variance was less than 1.5 times the RSD, were excluded.

Any deviations from these parameters are outlined in the individual assay summery documents.

Appendix

A.1 Format of Data Provided

For each assay, aligned, or feature extracted data has been provided as an $m \times n$ matrix of intensity values, accompanied by; a $m \times o$ matrix of sample IDs and metadata, a $n \times p$ matrix of feature identifications.

Each assay is provided as a separate folder. Within each assay folder sample measurements are found in the file `*StudyName Matrix Assay*_intensityData.csv`, rows in this file correspond to the sample metadata (including the sample ID provided to the NPC, under the heading 'Sample ID') found in `*StudyName Matrix Assay*_sampleMetadata.csv` while columns correspond to the observed features, characterised in `*StudyName Matrix Assay*_featureMetadata.csv`.

A.2 Glossary

.raw	MS: Original format for UPLC-MS data generated on instrument.
Resonance	NMR: An observed signal or set of signals associated with a chemical entity.

Feature	MS: An observed chemical entity, defined as a detected peak, plus any other peaks observed to arise from isotopes or chemical adducts, denoted as a characteristic m/z and observed region time NMR: The response observed at a specific chemical shift, as denoted in ppm.
Line-width	NMR: The half-height line width in Hz, of a representative resonance.
Long-term Reference (LTR)	An NPC-internal reference sample, used to track cross-study analytical stability.
m	The number of assays in a dataset, including study-samples, and where relevant SR.
mzNLD	MS: Vendor-proprietary format for UPLC-MS data. Generated in the course of feature-extraction in Progenesis QI.
n	The number of observations in a dataset, whether representing discrete features as in UPLC-MS or samples of a continuous waveform as in NMR.
ppm	NMR: Chemical shift scale for NMR data, used to identify peak locations in a manner independent of magnetic field strength.
Study Reference (SR)	A pooled mixture of every study sample, used for quality control purposes.
Study Sample (SS)	The sample set provided for analysis.
m/z	MS: Mass to charge ratio, the measured mass of a detected ion, as a factor of its ionisation state.
Retention time	MS: The characteristic elution time for an observed feature under specific chromatographic conditions (in seconds unless otherwise specified)

A.3 Bibliography

Dona A.C, *et. al.*, "Precision high-throughput proton NMR spectroscopy of human urine, serum, and plasma for large-scale metabolic phenotyping." *Anal. Chem.* 86.19 (2014): 9887-9894.

Lewis, M.R., *et. al.*, "Development and Application of UPLC-ToF MS for Precision Large Scale Urinary Metabolic Phenotyping." *Anal. Chem.* (2016).

Sarafian, M. H. *et al.*, "Objective set of criteria for optimization of sample preparation procedures for ultra-high throughput untargeted blood plasma lipid profiling by ultra performance liquid chromatography-mass spectrometry." *Anal Chem.* 86.12 (2014): 5766–5774.

Zelena E, *et. al.*, "Development of a robust and repeatable UPLC-MS method for the long-term metabolomic study of human serum", *Anal. Chem.* 81.4 (2009): 1357-1364.

Generated with the nPYc Toolbox version 0.1.1a.