International NMR-based Environmental Metabolomics Intercomparison Exercise

Mark R. Viant[1,*], Dan Bearden[2,*], Jacob G. Bundy[3], Ian Burton[4], Timothy W. Collette[5], Drew R. Ekman[5], Vilnis Ezernieks[6], Tobias Karakach[4], Ching Yu Lin[7,§], Simone Rochfort[6], Jeffrey S. de Ropp[7], Quincy Teng[5], Ronald S. Tjeerdema[7], John Walter[4] and Huifeng Wu[1]

[1] *School of Biosciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK*

[2] *National Institute of Standards and Technology, Analytical Chemistry Division, Hollings Marine Laboratory, 331 Fort Johnson Road, Charleston, SC29412, USA*

[3] *Department of Biomolecular Medicine, Division of Surgery, Oncology, Reproductive Biology, and Anaesthetics (SORA), Faculty of Medicine, Sir Alexander Fleming Building, Imperial College London, London, SW7 2AZ, UK*

[4] *NRC Institute for Marine Biosciences, 1411 Oxford Street, Halifax NS, B3H 3Z1, Canada*

[5] *U.S. Environmental Protection Agency, National Exposure Research Laboratory, 960 College Station Road, Athens, GA30605, USA*

[6] *Biosciences Research Division, Department of Primary Industries, 621 Sneydes Road, Werribee, Vic 3030, Australia*

[7] *Department of Environmental Toxicology, University of California, Davis, CA95616, USA*

[§]Current address: Institute of Environmental Health, College of Public Health, National Taiwan University, Taipei 100, Taiwan

*Equal contributions/corresponding authors

Corresponding authors:

Dr. Mark R. Viant

School of Biosciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK

Phone: +44 (0)121 414 2219

FAX: +44 (0)121 414 5925

Email: M.Viant@bham.ac.uk

Dr. Dan Bearden

National Institute of Standards and Technology, Hollings Marine Laboratory, 331 Fort Johnson Road, Charleston, SC 29412, USA

Phone: +1 843 762 8865

FAX: +1 843 762 8737

Email: Dan.Bearden@nist.gov

Abstract

Several fundamental requirements must be met so that NMR-based metabolomics and the related technique of metabonomics can be formally adopted into environmental monitoring and chemical risk assessment. Here we report an intercomparison exercise which has evaluated the effectiveness of ^{1}H NMR metabolomics to generate comparable datasets from environmentally derived samples. It focuses on laboratory practice that follows sample collection and metabolite extraction, specifically the final stages of sample preparation, NMR data collection (500, 600 and 800 MHz), data processing and multivariate analysis. Seven laboratories have participated from the USA, Canada, UK and Australia, generating a total of ten datasets. Phase 1 comprised the analysis of synthetic metabolite mixtures, while Phase 2 investigated European flounder (*Platichthys flesus*) liver extracts from clean and contaminated sites. Overall, the comparability of datasets from the participating laboratories was good. Principal components analyses (PCA) of the individual datasets yielded ten highly similar scores plots for the synthetic mixtures, with a comparable result for the liver extracts. Furthermore, the same metabolic biomarkers that discriminated fish from clean and contaminated sites were discovered by all the laboratories. PCA of the combined datasets showed excellent clustering of the multiple analyses. These results demonstrate that NMR-based metabolomics can generate data that are sufficiently comparable between laboratories to support its continued evaluation for regulatory environmental studies.

Introduction

Environmental metabolomics is an approach for investigating the interactions of living organisms with their environment. It complements other "-omic" methods such as transcriptomics and proteomics by characterizing the plethora of low molecular weight endogenous metabolites in a biological sample (*1-4*). Metabolomics and the related technique of metabonomics have considerable potential as a tool for environmental risk assessment of chemicals as well as for environmental monitoring and diagnostics (*5-10*). Previously it has been applied to both the terrestrial (*11, 12*) and aquatic environments (*13-18*) to investigate the effects of toxicants, disease and other environmental stressors on organism health and metabolism. As with any new approach it is critical that a thorough evaluation of experimental comparability and precision is conducted (see reference *19* for definition of terms). This is particularly true for fingerprinting ("-omic") technologies that attempt to measure hundreds or thousands of parameters simultaneously. Only then can the approach be considered for incorporation into risk assessment and monitoring programs (*5*). Here we report such an interlaboratory comparability study to evaluate one-dimensional ^{1}H nuclear magnetic resonance (NMR) spectroscopy-based metabolomics involving seven laboratories across North America, Europe and Australia.

Many factors are involved in conducting high-quality NMR metabolomic studies. The most important is good experimental design and execution, whether the study involves field collections or laboratory based experiments. A typical study involves the collection of biological material, extraction and measurement of metabolites using NMR spectroscopy or mass spectrometry, spectral processing and multivariate analysis. Typically the goal is to discover biomarkers that discriminate between different groups of samples (e.g., control versus stressed).

Previously, Keun et al. (*20*) assessed the reproducibility of an NMR metabolomics experiment by analyzing two identical sets of rat urine from an acute toxicity study. The analyses were performed at two sites and the resulting datasets were extremely similar when analyzed by principal components analysis (PCA), giving nearly identical descriptions of the metabolic responses to hydrazine. More recently, identical sets of human urine before and after dietary intervention were measured using 250, 400, 500 and 800 MHz NMR spectrometers (*21*),. When analyzed by partial least squares discriminant analysis (PLS-DA), the loadings were found to comprise of the same spectral regions implying that the same metabolites were discriminating pre- and post-dietary intervention, independent of magnetic field strength.

The primary purpose of our intercomparison exercise was to demonstrate the efficacy of NMR metabolomics measurements for environmental research using methods that are standard practice in the participating laboratories. It expands upon the earlier intercomparison studies by increasing the participating laboratories to seven, analyzing an environmentally relevant sample requiring more complex preparation than a biofluid such as urine (specifically liver extracts derived from environmentally sampled fish), and using a more diverse array of NMR spectrometers (Bruker and Varian instruments at 500, 600 and 800 MHz) and software packages. Additionally, it provides considerable focus on the ability of the approach to discover specific biomarkers that differentiate control from pollutant-exposed fish. Several specific objectives were identified: (i) to begin to develop best-practice protocols for an annual or bi-annual comparability improvement program incorporating the wider community; (ii) to demonstrate the capability of the seven laboratories to conduct these measurements and to ask the questions: can independent laboratories discriminate metabolic fingerprints of ‘control’ versus ‘stressed’

phenotypes, and can they discover the same molecular biomarkers?; and (iii) to raise awareness of the field and confidence in NMR-based environmental metabolomics.

The exercise has several key features. First, it focuses on issues related to standard laboratory practice that come after biological sample collection and metabolite extraction, so samples were prepared in a single laboratory to provide consistency. Secondly, a guidance document was developed for the participating laboratories that provided sufficient detail to allow qualitative and quantitative comparison of the results, but enough flexibility to allow laboratory practices to vary. Thirdly, reporting of results was conducted anonymously by having a single data coordinator remove laboratory identifications prior to analysis. Finally, it was decided that both 'simple' (chemically defined synthetic mixtures) and 'complex' (tissue extracts) sample sets would be analyzed that reflect some of the complexities of environmental metabolomics research. The latter consisted of fish liver extracts from European flounder (*Platichthys flesus*) sampled from the mouths of an unpolluted and a contaminated river in the UK.

Materials and Methods

Synthetic mixtures

Six mixtures of metabolites (S1-S6) were prepared by one laboratory with various concentrations of glucose, citrate, fumurate, glutamine, alanine and nicotinate (Sigma-Aldrich, UK). Sample S1 was split into six fractions (S1a-S1f) for assessing repeatability. The metabolites and their concentrations were chosen to test the strength of the NMR methods. Specifically they were

common metabolites from different chemical classes at realistic concentrations (ca. 45 μmol/L to 10 mmol/L, or zero), their NMR resonances spanned a reasonable chemical shift with some overlap of different compounds, and one metabolite was selected with strong pH dependence. The bulk mixtures were prepared in 100 mmol/L sodium phosphate buffer (pH 7.0), quantitatively transferred to microcentrifuge tubes, dried in a centrifugal concentrator and sealed with parafilm. A buffer blank was also included. The preparation laboratory analyzed one batch of samples to ensure there was no contamination. This was a blind run, with the instrument operator not knowing the details of the mixture. Information concerning integration and exclusion regions was assessed from this dataset (described below). Then each participating laboratory was sent an identical batch of dried samples at ambient temperature.

Tissue extracts

Adult female European flounder (*Platichthys flesus*) were collected from the mouths of the Rivers Alde (unpolluted control site) and Tyne (polluted site) in the UK. After the fish were sacrificed, liver tissues were immediately dissected and snap-frozen in liquid nitrogen, and stored at -80°C until extraction by one laboratory. Six "exposed" liver samples (BE1-BE6; each comprised of two randomly selected, pooled livers, ca. 400 mg) and six "control" liver samples (BC1-BC6; where BC1 consisted of four randomly selected, pooled livers, ca. 1200 mg, and BC2-BC6 each comprised of two pooled livers, ca. 400 mg) were each extracted using a methanol:chloroform:water method and Precellys-24 bead-based homogenizer (Stretton Scientific Ltd, UK), as described previously (*22*, *23*). Sample BC1, after extraction, was split into three fractions (BC1a-BC1c) for assessing repeatability. All the polar metabolite fractions in aqueous methanol were dried, resuspended in 100 mmol/L sodium phosphate buffer (pH 7.0),

quantitatively transferred to microcentrifuge tubes, dried again, and sealed with parafilm. A buffer blank was also included. Each participating laboratory was sent an identical batch of dried samples at ambient temperature. Information concerning integration and exclusion regions was determined by the coordinator laboratory (described below).

^{1}H NMR spectroscopy

Immediately prior to NMR analysis, analysts in each of the seven laboratories resuspended the dried samples in 99.9 atom % D_2O containing 1.0 mmol/L sodium 3-trimethylsilyl-2,2,3,3-d_4-propionate (TSP) internal standard, according to Protocol 1 (supporting information, SI). Samples were then analyzed using ten NMR spectrometers across the seven laboratories (Table S1), including both Bruker and Varian instruments at 500, 600 and 800 MHz. One-dimensional (1-D) ^{1}H NMR spectra were obtained using presaturation to suppress the water resonance and recorded as described in Protocol 2 (SI). Post-acquisition processing was conducted on every spectrum independently by each laboratory, including zero filling, apodization, Fourier transformation, manual phasing and baseline correction, and calibration, as described in Protocol 3 (SI). Although laboratories were asked to report the hardware and software used, no attempt was made to standardize on one platform. Finally, ^{1}H-^{13}C HSQC and HMBC NMR experiments were conducted on a single liver extract by a coordinating laboratory to assist with metabolite identification.

Data processing and analysis

NMR spectra of the synthetic mixtures and tissue extracts were analyzed using multiple methods. First, defined spectral regions were integrated by each participating laboratory according to

Protocol 4 (SI), including the TSP internal standard. These peak areas were normalized and the resulting metabolite concentrations were used to calculate the quality metrics described below. Next, each participating laboratory converted its NMR spectra to a format for multivariate analysis by segmentation into 0.005-ppm bins, exclusion of the spectral region surrounding the residual suppressed water resonance, and normalization to TSP (synthetic mixtures) or to a total spectral area of one (tissue extracts), according to Protocol 5 (SI). The binned spectral data were mean centered and analyzed by PCA, again using several different software packages (Table S1). Analyses conducted by each laboratory are referred to as "individual multivariate analyses". The results from all 10 analyses from the 7 laboratories, including the integrated peak areas, binned spectra, PCA scores and loadings were reported to the data coordinator who removed all identifying information. All the loadings data were compared to assess the degree of consistency of discovering biomarkers, described below. Finally, PCAs were conducted on each of the entire synthetic mixture and tissue extract datasets (referred to as "combined multivariate analyses").

Results and Discussion

Analysis of synthetic mixtures

The analysis of the synthetic mixtures was designed to address (1) sample handling (re-hydrating in D_2O, adding internal standard); (2) NMR acquisition (shimming, pulse-width determination, stable spectrometer operation, sensitivity and dynamic range); (3) post-acquisition processing (apodization, phasing, baseline correction, calibration, peak integration); and (4) spectral

processing (binning, exclusion regions, normalization) and PCA. Representative NMR spectra of synthetic sample S1a show many peaks arising from the 6 pure metabolites (Fig. 1a).

Comparability and precision of integrated peak areas in synthetic mixtures

To gauge the semi-quantitative capability of the participating laboratories, integration of six selected metabolite peaks plus the TSP peak was performed by the individual laboratories on their data (Table S2). By analyzing the six sample replicates for each laboratory (S1a-S1f) normalized to TSP, estimates of the individual laboratory repeatability were made, and by analyzing all of the samples across laboratories, inter-laboratory comparability and intermediate precision were assessed (*24, 25*). Quality metrics were based on a set of 'exercise consensus values' for the mean and standard deviation of each integrated peak after an outlier analysis. The outlier analysis is based on robust statistics and draws from ideas related to Mandel's h and k graphical tests (*26*). The analysis was performed by iteratively constructing box-plots of the grouped data and eliminating data that fell outside the box-plot whiskers (representing the upper or lower quartile ± 1.5 times the interquartile range) until, after 2 or 3 iterations, all the remaining data was within the box-plot whiskers. Of the total of 110 NMR spectra of synthetic mixtures (11 spectra for each of 10 datasets) comprising 770 semi-quantitative values, only 35 individual values were determined to be outliers, including four spectra (S1a and S1b in datasets 00115 and 00122, all from the same laboratory and for which all 6 metabolites per spectrum were outliers), for which a simple, systematic error was detected. These 4 spectra were excluded from all subsequent analyses.

The precision for each of the six analytes of each replicate sample (S1a-S1f) was given a p-score compared to a target percent relative standard deviation (%RSD) of 15% (*27*). This score is denoted "p-score(15%)" and is computed as:

$$p-score(15\%)=\frac{\%RSD(analysis)}{\%RSD(\text{target})},\quad with\ \%RSD(\text{target})=15\%$$

The ranges of p-score(15%) were evaluated as "Comparable" for $|p|\leq 2$, "Uncertain" for $2<|p|\leq 3$, and "Problematic" for $|p|>3$. In addition, each semi-quantitative peak (66 per dataset arising from 11 spectra (S1a-S1f, S2-S6) of 6 analytes) was given two z-scores based on two different criteria; one compared to 25% of the exercise assigned value and one compared to the exercise assigned standard deviation (*24*). These are denoted "z-score(25%)" and "z-score(s)". The z-scores are computed as:

$$z-score=\frac{(x-\overline{X})}{\sigma}$$

where $\overline{X}$ is the consensus exercise assigned mean, and

$\sigma = 25\%\ of\ the\ exercise\ assigned\ mean\ \ for\ z-score(25\%)$

or $\sigma = the\ exercise\ assigned$ standard $deviation\ \ for\ z-score(s)$.

The ranges of z-scores were evaluated as "Comparable", "Uncertain" or "Problematic" analogously to the p-score(15%). The repeatability performance of the participating laboratories is excellent, with the p-score(15%) universally Comparable, reflecting the highly repeatable nature of NMR data acquisition (Table 1). The comparability performance was very good both when the target comparability was 25% of the exercise assigned value (z-score(25%)), and when the target comparability was assessed by the z-score(s) values. One sample (S3) was prepared with peak 3 (citrate) at zero concentration, but the resulting assigned value for the exercise was

not zero, and each dataset was flagged with a "Problematic" z-score(25%) score while all but one dataset received "Comparable" z-score(s) ratings for this measure.

Individual multivariate analyses of synthetic mixtures

Each laboratory conducted PCA on their NMR data using Protocol 5 and their choice of software (Table S1). Each scores plot reflects the "metabolic" differences between the six samples comprising one replicate of S1 as well as S2-S6, taking into account the entire chemical shift range of each binned NMR spectrum. Collectively, the 10 scores plots enable a visual comparison between all the synthetic mixture datasets (Fig. S1a-j) which show a remarkably consistent pattern. Table S4 further highlights the similarity between these 10 PCA models by summarizing the variances captured by principal components (PCs) 1-5, with PC1 ranging from 40.4-45.9% and PC2 from 35.0-37.4%. Note that this includes models based on 500, 600 and 800 MHz NMR data. The 10 associated PC1 loadings plots were used to visually assess if individual laboratories could reveal which metabolites discriminated the synthetic mixtures. Fig. S2 shows just those bins corresponding to the 20 largest PC1 loadings for each PCA. Most of the 10 independent analyses found the same pattern of discriminatory compounds, although some analyses were clearly different with this limited set of top-20 loadings (further interpretation in Fig. S2 caption).

Combined multivariate analyses of synthetic mixtures

Three further PCAs were conducted after the binned data supplied by the participants were combined (total of 110 spectra at 500, 600 and 800 MHz, with 4 outlier spectra removed as discussed above). The aim was to assess the similarity of the binned synthetic mixture spectra

across laboratories. Prior to PCA, the 11 spectra from each laboratory were multiplied by a single scaling-factor such that the total spectral area of the S1 replicate was set at unity, enabling comparison across laboratories. The first PCA was conducted on the unique samples (one replicate of S1 as well as S2-S6, for all 10 datasets). The scores plot shows a high degree of similarity across the 10 analyses, which is particularly surprising considering that this includes data from different spectrometer frequencies (Fig. 2). Furthermore, the overall appearance is similar to the 10 PCA scores plots from the individual analyses (Fig. S1a-j). To assess the comparability of the replicates (S1a-S1f) from each laboratory compared to the relatively large metabolic differences between samples, the PCA model of unique samples (S1-S6) calculated above was applied to samples S1b-S1f. The scores plot clearly highlights the comparability of the technical replicates from each laboratory (Fig. S3). The third PCA was conducted on replicate samples only (S1a-S1f). Not unexpectedly, the scores plot reveals a dependence on spectrometer frequency which is most apparent along PC1, while variation at any one frequency tends to occur along PC2 (Fig. S4).

Analysis of biological tissue extracts

Analysis of the tissue extracts was designed to address the same factors as for the synthetic mixtures, but also to include real-world biological variation to demonstrate consistency of sample classification and biomarker discovery across laboratories. Representative NMR spectra of fish liver extract BC1a show hundreds of peaks arising from the extracted liver metabolome (Fig. 1b).

Comparability and precision of integrated peak areas in tissue extracts

Each laboratory integrated eleven selected spectral regions plus TSP (Table S3). Data were subjected to outlier analysis as described above, then individual laboratory repeatability was evaluated using the three tissue extract replicates (BC1a-BC1c), and inter-laboratory comparability and intermediate precision were assessed for all the samples across laboratories, as described for the synthetic mixture data. For the biological samples, the precision scores are very good at the p-score(15%) level (Table 1). For the outlier analysis, only 86 of the 1540 individual integral values were determined to be outliers. The z-score(25%) scores indicate good performance, as does the z-score(s) assessment (Table 1). Judging from the number of "Problematic" results for the z-score(s) assessment (median=10), this is a more stringent assessment than z-score(25%) (median=0) for the biological samples. There are more "Problematic" results for biological compared to synthetic samples. This could represent poor biological sample homogeneity, but given the bulk extraction protocol, this seems unlikely. The patterns in the outlier analysis indicate systematic problems with a single or a small number of integral regions across all samples for some datasets, and this could result from systematic differences between laboratories such as baseline or phase correction issues or sample temperature differences (which could cause small peak shifts). Two datasets from one laboratory (09541 and 07042) initially contained systematic errors related to baseline correction and other issues; the data were corrected by the submitting laboratory before being used in subsequent analyses.

Individual multivariate analyses of tissue extracts

Each laboratory conducted a PCA of the NMR metabolite fingerprints of the 12 unique liver extracts BC1a, BC2-BC6 and BE1-BE6. Collectively, the 10 PCA scores plots from the seven

laboratories enabled a visual comparison between all the tissue extract datasets (Fig. S5a-j). As for the synthetic mixture results, the 10 scores plots of the biological samples were similar across all laboratories. For 8 of the analyses the PC1 axis was highly significant in discriminating between fish from control and polluted sites, with p-values ranging from 0.002 to 0.005 (Table S5), highlighting the similarity between the independent studies. For the remaining two datasets, the separation on PC1 was less significant ($p=0.033$ for dataset 00115) or near significant ($p=0.081$ for dataset 00122, whereas separation along PC2 was highly significant, $p=0.007$). Table S5 further compares the 10 PCA models by summarizing the percent variances captured by PCs 1-5. Not surprisingly these values have wider ranges than for the synthetic mixtures, with variance explained by PC1 from 39.6-50.7% and PC2 from 14.9-20.7%. Since a primary objective of many metabolomics studies is biomarker discovery, the loadings data from the 10 PCAs were analyzed to determine if individual laboratories discovered the same metabolic differences between fish from the control versus polluted site. Visual comparison of the bins corresponding to the 20 largest PC1 loadings from each PCA suggests a high degree of consistency (Fig. S6).

A more extensive statistical comparison of the loadings was then undertaken. First, for each individual PCA performed by the participants, we located the signal intensities for the 20 bins with the largest PC1 loadings. A t-test was performed on each of these bins to determine if the signal intensities could discriminate the control and polluted sites (Table 2). Eight of the ten analyses discovered between 10-14 significant bins, with two analyses (00115 and 00122, by one laboratory) discovering only 5-7 significant bins. Non-significant bins ($p\geq0.05$) were discarded. Considering only the top-20 loadings, biomarker discovery was extremely consistent across the

participating laboratories: e.g., both peaks associated with unknown metabolite #1 were significantly increased in exposed fish in all ten analyses (by 1.46 to 1.90-fold); glucose significantly decreased in all ten analyses (by 1.32 to 2.32-fold); lactate significantly decreased in nine analyses (by 1.77 to 2.31-fold); and unknown metabolite #2 significantly increased in eight analyses (by 1.70 to 2.13-fold). These metabolic differences between fish from the Rivers Alde and Tyne could be due to several factors, e.g. genetic differences, diet or water quality.

There are more powerful methods than PCA for biomarker discovery (*28*). PCA is an unsupervised approach, and there is no *a priori* reason why sample separation should occur along a specific PC axis. For example, Fig. S5 clearly shows that maximum discrimination between sites within the first two components would be obtained by a combination of PC1 and PC2 scores. In addition, PCA tends to privilege changes in high-concentration metabolites, as these possess high variance. Nonetheless, we chose to use PCA as an example of a common, robust multivariate technique frequently used in metabolomics studies; its unsupervised nature is positively beneficial in this context in explicitly showing the full effects of potentially confounding (e.g. between-laboratory) variance. Thus, although many other bins outside of the top-20 loadings would be significantly different between control and polluted sites, and a more active data-mining approach would identify more potential biomarkers, this was not our current aim. Overall these results demonstrate a high confidence in biomarker discovery by NMR metabolomics. It is also worth noting that the two 800 MHz datasets discovered more significant biomarkers (in the top-20 loadings) than the 500 and 600 MHz datasets.

Combined multivariate analyses of tissue extracts

Metabolic differences between fish, both intra- and inter-group, were subtle compared to differences between the synthetic mixtures. Consequently, analyses of the tissue extract NMR datasets had to be conducted for each spectrometer frequency, since a PCA of all datasets combined revealed a strong dependency on spectrometer frequency (Fig. S8) as one would expect. This dependency arises because the spin-spin coupling patterns at 500, 600 and 800 MHz are not identical along the binned chemical shift axis. Of the 120 spectra (6 control and 6 exposed per dataset), four were removed from all PCAs because three or more of their integrated spectral regions were outliers (described above); these comprised BC5 (dataset 02861), BC6 (00258) and BE3 (00115 and 00122). Separation of control and polluted sites as well as high inter-laboratory measurement repeatability is clearly evident from the scores plot of the 500 MHz data (Fig. 3a). The sites are also separated along PC1 for the 600 MHz data, but one dataset was separated from the other three along PC2 (Fig. 3b). For the 800 MHz data, sites were separated along PC2, with differences between the two datasets dominating PC1 (Fig. S9). Inspection of the corresponding PC1 and PC2 loadings plots (Fig. S10) reveals that metabolic differences between control and polluted fish are highly reproducible across spectrometer frequencies, and that the primary source of inter-laboratory differences at any one spectrometer frequency is peak shifting. Shifting can arise for several reasons, but considering that all laboratories analyzed identical samples (same pH, same divalent metal ion content), it is highly likely to have been caused by differences in sample temperature during NMR acquisition. A PCA of the entire 500 MHz dataset, including the technical replicates BC1a-BC1c, again highlights the high comparability of the NMR measurements (Fig. S11).

Recommendations from intercomparison exercise

Here we have demonstrated that the comparability and precision of the laboratories participating in this exercise are good, reflecting the relative ease of obtaining good semi-quantitative results from NMR spectra. The precision of the measurements is very high, but the comparability can vary in subtle systematic ways that warrant further attempts to standardize the practice, for example through a further comparability improvement program. Furthermore we have demonstrated that the collection and statistical analysis of NMR metabolomics spectra of both synthetic and biological samples, given reasonably well defined protocols, can be conducted in a consistent manner by seven independent laboratories, yielding similar multivariate classification models and associated biomarkers. Based upon these analyses we recommend that improving the consistency of sample temperature in the NMR spectrometer, including temperature calibration of the probe (*29*), may improve the measurement repeatability by decreasing peak shifting. We further recommend that this exercise be repeated, with the inclusion of additional laboratories, and that it be expanded in scope to include some basic (e.g., greater specification of baseline subtraction methods) and advanced protocols (e.g., 2-dimensional J-resolved NMR spectroscopy (*30, 31*) to circumvent the differences in spin-spin coupling patterns that arises in multi-frequency NMR studies). The results from this first exercise lend considerable support to the high quality of NMR metabolomics data and its continued evaluation for incorporation into regulatory environmental science.

Acknowledgments

This work was partially funded by the UK NERC via an International Opportunities Award (NE/D000793/1) and Post Genomic and Proteomic Directed Program (NE/C507661/1). MRV additionally thanks the NERC for an Advanced Fellowship (NER/J/S/2002/00618). The authors

400 thank Dr. Brett Lyons (Centre for the Environment, Fisheries and Aquaculture Science) for the
401 flounder livers. We acknowledge the support of the Hollings Marine Laboratory NMR Facility
402 and David Keiser (Bio21) for assistance and access to NMR instrumentation.
403

404 **Disclaimer**

405 Commercial equipment or materials are identified in this paper to specify adequately the
406 experimental procedure. Such identification does not imply recommendation or endorsement by
407 NIST or EPA, nor does it imply that the materials or equipment identified are necessarily the
408 best available for the purpose. This paper has been reviewed in accordance with the U.S. EPA's
409 peer and administrative review policies and approved for publication.
410

411 **Supporting Information Available**

412 The NMR datasets are freely available from the corresponding authors for re-analysis.
413 Supporting information documents the five protocols developed for the intercomparison exercise
414 and describes further analyses of the synthetic and biological datasets. This information is
415 available free of charge via the Internet at http://pubs.acs.org.
416

Literature Cited

1. Kell, D. B., Metabolomics and systems biology: making sense of the soup. *Curr. Opin. Microbiol.* **2004,** *7*, 296-307.
2. Lindon, J. C.; Holmes, E.; Bollard, M. E.; Stanley, E. G.; Nicholson, J. K., Metabonomics technologies and their applications in physiological monitoring, drug safety assessment and disease diagnosis. *Biomarkers* **2004,** *9*, 1-31.
3. Pelczer, I., High-resolution NMR for metabomics. *Curr. Opin. Drug Discov. Devel.* **2005,** *8*, 127-133.
4. Robertson, D. G., Metabonomics in toxicology: A review. *Toxicol. Sci.* **2005,** *85*, 809-822.
5. Genomics Task Force Workgroup, *Potential Implications of Genomics for Regulatory and Risk Assessment Applications at EPA,* U.S. Environmental Protection Agency (EPA 100/B-04/002), **2004**.
6. Rochfort, S., Metabolomics reviewed: A new "Omics" platform technology for systems biology and implications for natural products research. *J. Nat. Prod.* **2005,** *68*, 1813-1820.
7. Lin, C. Y.; Viant, M. R.; Tjeerdema, R. S., Metabolomics: Methodologies and applications in the environmental sciences. *J. Pest. Sci.* **2006,** *31*, 245-251.
8. Viant, M. R., Metabolomics of aquatic organisms: the new 'omics' on the block. *Mar. Ecol.-Prog. Ser.* **2007,** *332*, 301-306.
9. Samuelsson, L. M.; Larsson, D. G. J., Contributions from metabolomics to fish research. *Mol. BioSys.* **2008,** *(accepted)*.
10. Viant, M. R., Recent developments in environmental metabolomics. *Mol. BioSys.* **2008,** *(accepted)*.

11. Bundy, J. G.; Spurgeon, D. J.; Svendsen, C.; Hankard, P. K.; Weeks, J. M.; Osborn, D.; Lindon, J. C.; Nicholson, J. K., Environmental metabonomics: Applying combination biomarker analysis in earthworms at a metal contaminated site. *Ecotoxicology* **2004,** *13*, 797-806.

12. Bundy, J. G.; Keun, H. C.; Sidhu, J. K.; Spurgeon, D. J.; Svendsen, C.; Kille, P.; Morgan, A. J., Metabolic profile biomarkers of metal contamination in a sentinel terrestrial species are applicable across multiple sites. *Environ. Sci. Technol.* **2007,** *41*, 4458-4464.

13. Viant, M. R.; Rosenblum, E. S.; Tjeerdema, R. S., NMR-based metabolomics: A powerful approach for characterizing the effects of environmental stressors on organism health. *Environ. Sci. Technol.* **2003,** *37*, 4982-4989.

14. Solanky, K. S.; Burton, I. W.; MacKinnon, S. L.; Walter, J. A.; Dacanay, A., Metabolic changes in Atlantic salmon exposed to Aeromonas salmonicida detected by ^{1}H-nuclear magnetic resonance spectroscopy of plasma. *Dis. Aquat. Org.* **2005,** *65*, 107-114.

15. Rosenblum, E. S.; Tjeerdema, R. S.; Viant, M. R., Effects of Temperature on Host-Pathogen-Drug Interactions in Red Abalone, *Haliotis rufescens*, determined by ^{1}H NMR Metabolomics. *Environ. Sci. Technol.* **2006,** *40*, 7077-7084.

16. Hines, A.; Oladiran, G. S.; Bignell, J. P.; Stentiford, G. D.; Viant, M. R., Direct sampling of organisms from the field and knowledge of their phenotype: Key recommendations for environmental metabolomics. *Environ. Sci. Technol.* **2007,** *41*, 3375-3381.

17. Ekman, D. R.; Teng, Q.; Jensen, K. M.; Martinovic, D.; Villeneuve, D. L.; Ankley, G. T.; Collette, T. W., NMR analysis of male fathead minnow urinary metabolites: A potential approach for studying impacts of chemical exposures. *Aquat. Toxicol.* **2007,** *85*, 104-112.

18. Ekman, D. R.; Teng, Q.; Villeneuve, D. L.; Kahl, M. D.; Jensen, K. M.; Durhan, E. J.; Ankley, G. T.; Collette, T. W., Investigating compensation and recovery of fathead minnow (Pimephales promelas) exposed to 17 alpha-ethynylestradiol with metabolite profiling. *Environ. Sci. Technol.* **2008,** *42*, 4188-4194.

19. Duewer, D. L.; Kline, M. C.; Sharpless, K. E.; Thomas, J. B.; Gary, K. T., Micronutrients measurement quality assurance program: Helping participants use interlaboratory comparison exercise results to improve their long-term measurement performance. *Anal. Chem.* **1999,** *71*, 1870-1878.

20. Keun, H. C.; Ebbels, T. M. D.; Antti, H.; Bollard, M. E.; Beckonert, O.; Schlotterbeck, G.; Senn, H.; Niederhauser, U.; Holmes, E.; Lindon, J. C.; Nicholson, J. K., Analytical reproducibility in ^{1}H NMR-based metabonomic urinalysis. *Chem. Res. Toxicol.* **2002,** *15*, 1380-1386.

21. Bertram, H. C.; Malmendal, A.; Petersen, B. O.; Madsen, J. C.; Pedersen, H.; Nielsen, N. C.; Hoppe, C.; Molgaard, C.; Michaelsen, K. F.; Duus, J. O., Effect of magnetic field strength on NMR-Based metabonomic human urine data comparative study of 250, 400, 500, and 800 MHz. *Anal. Chem.* **2007,** *79*, 7110-7115.

22. Lin, C. Y.; Wu, H. F.; Tjeerdema, R. S.; Viant, M. R., Evaluation of metabolite extraction strategies from tissue samples using NMR metabolomics. *Metabolomics* **2007,** *3*, 55-67.

23. Wu, H.; Southam, A. D.; Hines, A.; Viant, M. R., High throughput tissue extraction protocol for NMR- and MS-based metabolomics. *Anal. Biochem.* **2008,** *372*, 204-212.

24. Thompson, M.; Ellison, S. L. R.; Wood, R., The International Harmonized Protocol for the proficiency testing of analytical chemistry laboratories (IUPAC technical report). *Pure Appl. Chem.* **2006,** *78*, 145-196.

25. Hibbert, D. B., *Quality Assurance for the Analytical Chemistry Laboratory*. Oxford University Press, **2007**, p 150.

26. *Accuracy (trueness and precision) of measurement methods and results - Practical guidance for the use of ISO 5725-2:1994 in designing, implementing and statistically analysing interlaboratory repeatability and reproducibility results.* First edition. (ISO/TR 22971:2005), **2005**.

27. NIST Accuracy (z-score) and Precision (p-score) Assessments. http://www.tdi-bi.com/analytical_services/nist/z-p-scores.htm (accessed 2008).

28. Rousseau, R.; Govaerts, B.; Verleysen, M.; Boulanger, B., Comparison of some chemometric tools for metabonomics biomarker identification. *Chemom. Intell. Lab. Sys.* **2008,** *91*, 54-66.

29. Findeisen, M.; Brand, T.; Berger, S., A ^{1}H-NMR thermometer suitable for cryoprobes. *Magn. Reson. Chem.* **2007,** *45*, 175-178.

30. Viant, M. R., Improved methods for the acquisition and interpretation of NMR metabolomic data. *Biochem. Biophys. Res. Comm* **2003,** *310*, 943-948.

31. Wang, Y. L.; Bollard, M. E.; Keun, H.; Antti, H.; Beckonert, O.; Ebbels, T. M.; Lindon, J. C.; Holmes, E.; Tang, H. R.; Nicholson, J. K., Spectral editing and pattern recognition methods applied to high-resolution magic-angle spinning H-1 nuclear magnetic resonance spectroscopy of liver tissues. *Anal. Biochem.* **2003,** *323*, 26-32.

Table 1. Summary of quantitation performance scores for the Synthetic and Biological datasets. Results are presented as triplets of values representing the number of **C**omparable, **U**ncertain, or **P**roblematic results based on the criterion at the head of each column. The z-scores are measures of comparability and the p-scores are measures of precision.

Dataset	Synthetic (C/U/P)			Biological (C/U/P)		
	z-Score (25%)	**z-Score (s)**	**p-Score (15%)**	**z-Score (25%)**	**z-Score (s)**	**p-Score (15%)**
00115	52/1/1	54/0/0/	6/0/0	154/0/0	134/7/13	11/0/0
00122	51/1/2	54/0/0	6/0/0	154/0/0	141/1/12	11/0/0
00258	64/0/2	62/3/1	6/0/0	137/16/1	129/5/20	11/0/0
00333	64/1/1	66/0/0	6/0/0	140/13/1	152/1/1	11/0/0
00711	64/0/2	64/0/2	6/0/0	152/2/0	153/1/0	11/0/0
00714	65/0/1	65/1/0	6/0/0	154/0/0	154/0/0	11/0/0
02861	62/1/3	63/2/1	6/0/0	152/2/0	146/1/7	11/0/0
07042	64/0/2	66/0/0	6/0/0	140/0/14	140/0/14	11/0/0
08865	65/0/1	66/0/0	6/0/0	154/0/0	154/0/0	11/0/0
09541	63/1/2	59/5/2	6/0/0	128/9/17	127/9/18	11/0/0
Median	*64/1/2*	*64/0/0*	*6/0/0*	*152/1/0*	*144/1/10*	*11/0/0*

Table 2. Metabolite intensity ratios between fish sampled from polluted (n=6) versus control (n=6) sites, and associated p-values (in brackets) from t-tests of the 12 intensities. These biomarkers that all significantly discriminate the two groups of fish were discovered by analyzing the top-20 PC1 loadings, which were calculated by the participants, from each of ten individual PCAs of samples BC1-BC6 and BE1-BE6.

Metabolite	Bin chemical shift (ppm)	Dataset (and NMR spectrometer frequency, MHz)									
		00115 (600)	**00122 (500)**	**00258 (600)**	**00333 (500)**	**00711 (600)**	**00714 (800)**	**02861 (500)**	**07042 (800)**	**08865 (500)**	**09541 (600)**
Lactate	1.33	-[1]	0.439 (0.0282)	0.565 (0.0091)	0.455 (0.0342)	0.550 (0.0106)	-	-	0.506 (0.0430)	-	-
	1.34	0.495 (0.0127)	0.480 (0.0181)	0.472 (0.0145)	0.490 (0.0212)	0.480 (0.0167)	0.529 (0.0150)	0.433 (0.0408)	0.521 (0.0286)	0.465 (0.0246)	-
unknown #1[2]	3.12	1.90 (0.0010)	1.86 (0.0011)	1.79 (0.0016)	1.89 (0.0007)	1.75 (0.0015)	1.88 (0.0006)	1.66 (0.0077)	1.83 (0.0044)	1.75 (0.0043)	1.82 (0.0016)
	3.31	1.69 (0.0028)	1.68 (0.0018)	1.61 (0.0005)	1.62 (0.0010)	1.62 (0.0024)	1.78 (0.0019)	1.46 (0.0179)	1.58 (0.0093)	1.84 (0.0004)	1.63 (0.0028)
unknown #2	3.14	2.00 (0.0001)	-	1.90 (0.0000)	1.93 (0.0001)	2.13 (0.0005)	2.13 (0.0002)	1.70 (0.0011)	-	1.99 (0.0001)	1.97 (0.0002)
unknown #3	3.28	-	-	-	-	-	1.20 (0.0256)	-	1.21 (0.0052)	1.20 (0.0321)	1.20 (0.0470)

Glucose	3.41	-	-	0.514 (0.0055)	-	0.548 (0.0015)	0.520 (0.0021)	-	0.572 (0.0047)	-	0.553 (0.0019)
	3.48	0.522 (0.0069)	0.527 (0.0070)		0.542 (0.0016)	0.504 (0.0014)	0.490 (0.0016)	0.464 (0.0011)	0.518 (0.0049)	0.511 (0.0025)	0.497 (0.0024)
	3.50	-	-	0.576 (0.0104)	0.540 (0.0013)	0.491 (0.0010)	0.490 (0.0017)	0.464 (0.0013)	-	0.526 (0.0050)	0.496 (0.0024)
	3.72	-	-	-	-	-	0.552 (0.0026)	-	0.579 (0.0059)	-	-
	3.73	-	-	0.569 (0.0044)	-	0.573 (0.0024)	-	-	0.595 (0.0051)	-	-
	3.84	-	-	0.584 (0.0040)	0.570 (0.0012)	0.548 (0.0018)	0.535 (0.0014)	0.506 (0.0007)	0.546 (0.0026)	0.554 (0.0012)	0.563 (0.0011)
	3.89	-	-	0.597 (0.0033)	0.587 (0.0012)	0.590 (0.0010)	0.547 (0.0020)	0.536 (0.0008)	0.552 (0.0036)	0.590 (0.0013)	0.576 (0.0010)
	3.91	-	-	-	0.628 (0.0007)	0.624 (0.0012)	0.607 (0.0026)	0.568 (0.0008)	0.674 (0.0078)	-	-
	4.65	0.470 (0.0030)	-	-	0.485 (0.0015)	-	0.468 (0.0016)	0.431 (0.0009)	0.714 (0.0180)	0.482 (0.0016)	-
	4.66	0.469 (0.0011)	-	-	0.485 (0.0016)	-	0.466 (0.0012)	0.437 (0.0010)	-	0.468 (0.0009)	0.759 (0.0106)

[1]Denotes that the particular bin in a dataset is not a significant bin with a top-20 PC1 loadings value. Of the 118 significant bins ($p<0.05$) with the 20 largest loadings from the 10 analyses, 5 only appear in single datasets and are not included above, and a further 8 have been removed when two bins contribute to the same peak.

[2]Peaks at 3.12 and 3.31 ppm are highly correlated (Fig. S7) and have been confirmed to arise from the same unknown metabolite via an HMBC NMR experiment (data not shown).

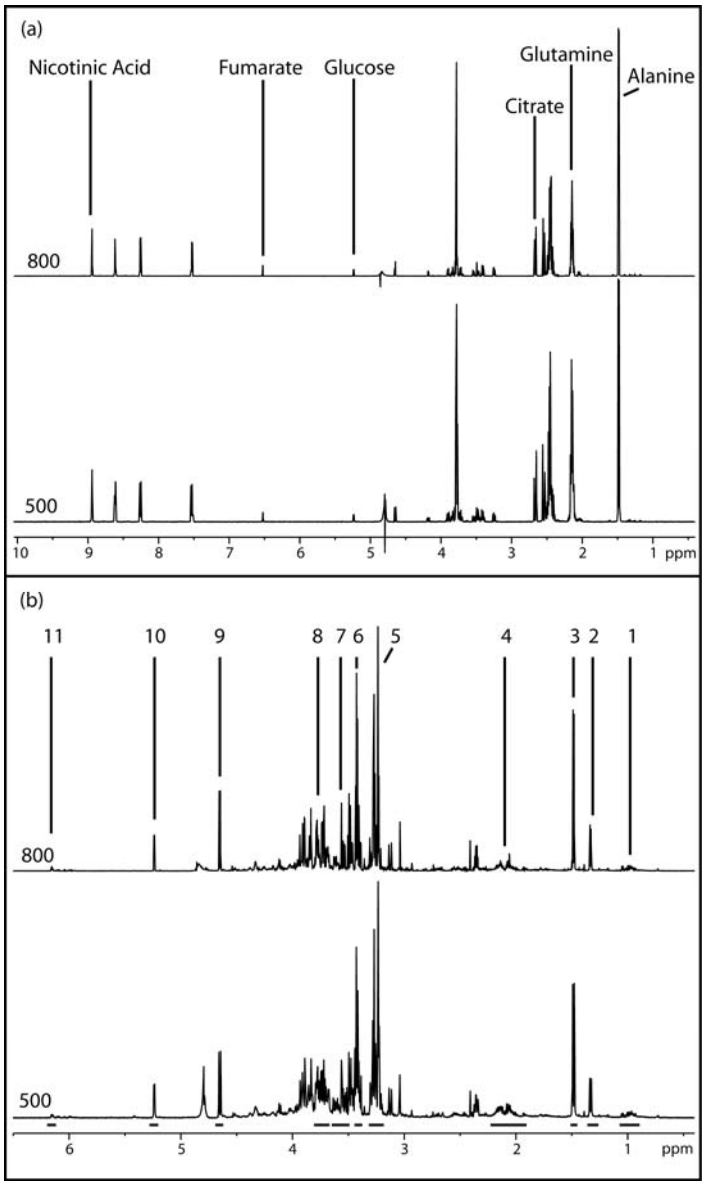
Figure Captions

Figure 1. Representative NMR spectra of the (a) synthetic and (b) fish liver extract samples at 500 MHz (lower) and 800 MHz (upper). In (b), the numbers above the peaks and the horizontal bars indicate the specific regions integrated for semi-quantitative analysis.

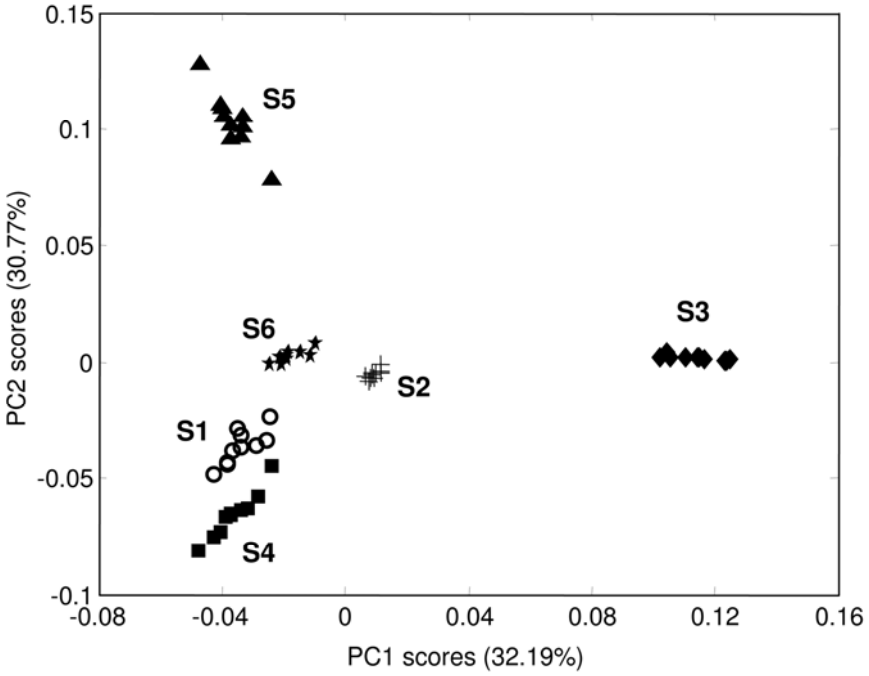
Figure 2. PCA scores plot of the combined 500, 600 and 800 MHz NMR spectra of the synthetic mixtures from all 10 analyses by the participating laboratories. Samples comprise a single replicate of S1 and the unique samples S2-S6, each n=10.

Figure 3. PCA scores plots from analysis of (a) four 500 MHz and (b) four 600 MHz NMR datasets of the liver extracts. For each dataset the samples comprise a single replicate of control BC1 and the unique controls BC2-BC6 (open symbols), and exposed BE1-BE6 (closed symbols). While all four 500 MHz datasets show high reproducibility and separation of control and exposed fish (solid lines drawn between mean scores for each sample, for the two different sites), this is only true for three of the four 600 MHz datasets (solid lines) with the fourth separated along PC2 (dotted lines). The outlying dataset (09541) was one of two from the same laboratory that had problems associated with spectral processing and peak shifting (see further details in Fig. S10c).

Figure 1

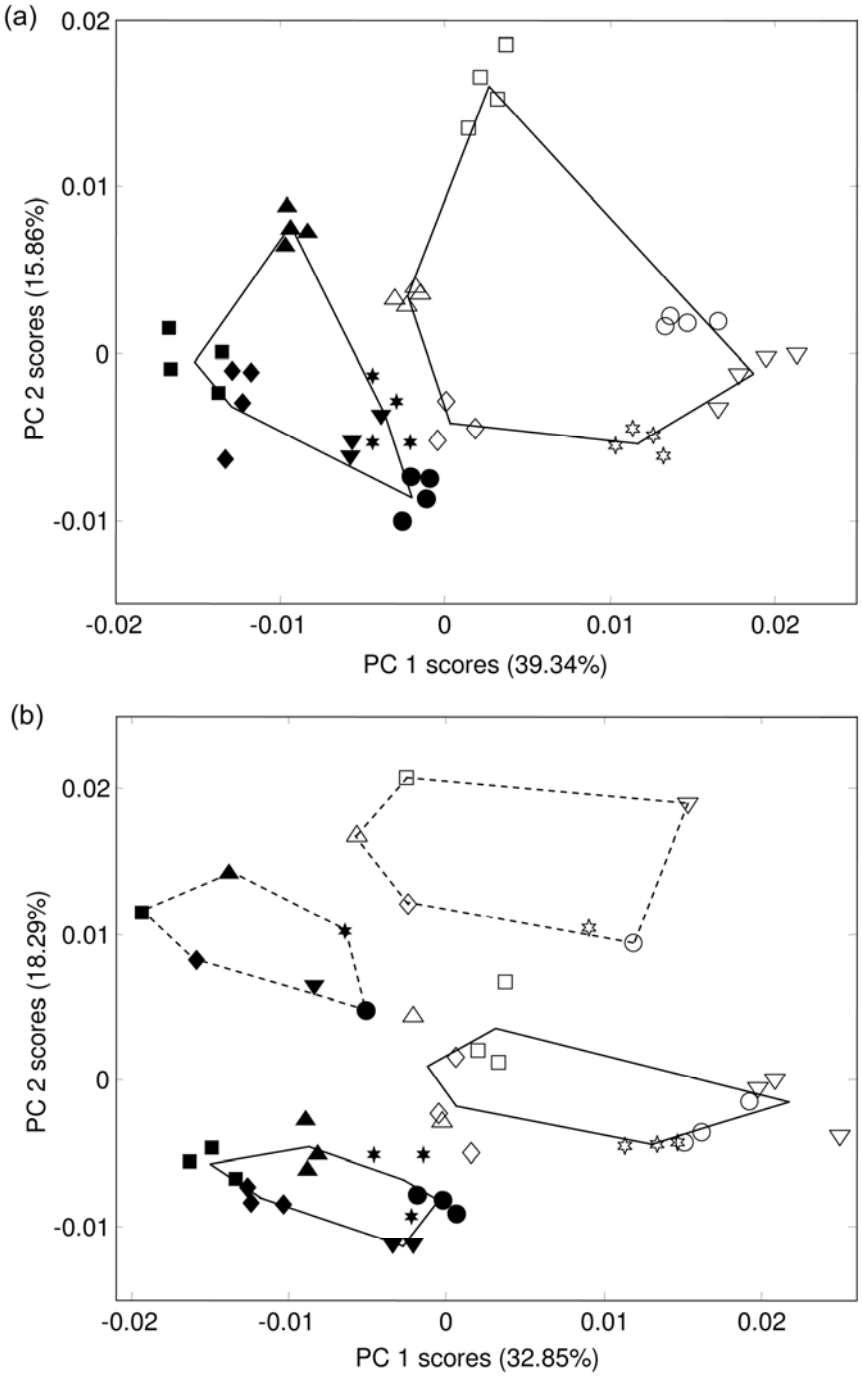

(a)
Nicotinic Acid
Fumarate
Glucose
Glutamine
Citrate
Alanine
800
500
10 9 8 7 6 5 4 3 2 1 ppm
(b)
11
10
9
8 7 6
5
4
3 2
1
800
500
6 5 4 3 2 1 ppm

546 Figure 2
547
548



549
550

551 Figure 3
552



553
554