Anal. Chem. **2009,** *81,* 7038–7046

Development and Performance of a Gas Chromatography−Time-of-Flight Mass Spectrometry Analysis for Large-Scale Nontargeted Metabolomic Studies of Human Serum

Paul Begley,*,[†] Sue Francis-McIntyre,*,[†] Warwick B. Dunn,[‡] David I. Broadhurst,[†] Antony Halsall,[†] Andy Tseng,[†] Joshua Knowles,[§] HUSERMET Consortium, Royston Goodacre,[‡] and Douglas B. Kell[†]

Bioanalytical Sciences Group and Manchester Centre for Integrative Systems Biology, School of Chemistry, and School of Computer Science, Manchester Interdisciplinary Biocentre, University of Manchester, M1 7DN, U.K.

A method for the preparation and GC−TOF-MS analysis of human serum samples has been developed and evaluated for application in long-term metabolomic studies. Serum samples were deproteinized using 3:1 methanol/serum, dried in a vacuum concentrator, and chemically derivatized in a two-stage process. Samples were analyzed by GC−TOF-MS with a 25 min analysis time. In addition, quality control (QC) samples were used to quantify process variability. Optimization of chemical derivatization was performed. Products were found to be stable for 30 h after derivatization. An assessment of within-day repeatability and within-week reproducibility demonstrates that excellent performance is observed with our developed method. Analyses were consistent over a 5 month period. Additional method testing, using spiked serum samples, showed the ability to define metabolite differences between samples from a population and samples spiked with metabolites standards. This methodology allows the continuous acquisition and application of data acquired over many months in long-term metabolomic studies, including the HUSERMET project (http://www.husermet.org/).

Metabolomics has been shown to be an informative method for investigating the genotype or phenotype[1,2] of biological systems. The profiling of metabolites in biological systems has been of interest for many decades, with the work of Williams et al.[3] cited as an early demonstration of “metabolic patterns” unique to individuals. Recent technological advances have increased the sensitivity and range of metabolites which can be detected concurrently. Modern metabolomics is characterized by a commitment to statistically informed study designs,[4−6] comprehensive biochemical and data analysis,[4] and defined reporting standards[7] allowing the robust assessment of data without biased prior selection of metabolic study targets. The analytical platforms most commonly used are gas or liquid chromatography−mass spectrometry[8,9] (GC/MS[10,11] or LC−MS[12−15]) and nuclear magnetic resonance (NMR) spectroscopy,[16] and relative strengths and weaknesses of these have been assessed.[17,18] In the absence of a universally applicable analytical platform, complementary data can

* To whom correspondence should be addressed. Paul Begley, e-mail paul.begley@manchester.ac.uk; Sue Francis-McIntyre, e-mail sue.mcintyre@manchester.ac.uk.

[†] Bioanalytical Sciences Group, School of Chemistry, Manchester Interdisciplinary Biocentre.

[‡] Manchester Centre for Integrative Systems Biology, School of Chemistry, Manchester Interdisciplinary Biocentre.

[§] School of Computer Science, Manchester Interdisciplinary Biocentre.

[1] Fiehn, O.; Kopka, J.; Dormann, P.; Altmann, T.; Trethewey, R. N.; Willmitzer, L. *Nat. Biotechnol.* **2000**, *18*, 1157–1161.

[2] Gieger, C.; Geistlinger, L.; Altmaier, E.; de Angelis, M. H.; Kronenberg, F.; Meitinger, T.; Mewes, H. W.; Wichmann, H. E.; Weinberger, K. M.; Adamski, J.; Illig, T.; Suhre, K. *PLoS Genetics* **2008**, *4*, e1000282.

[3] Williams, R. J. *Biochemical Institute Studies IV. Individual Metabolic Patterns and Human Disease: An Exploratory Study Using Predominantly Paper Chromatographic Methods*, U. Texas Publication No. 5109, University of Texas: Austin, TX, 1951.

[4] Broadhurst, D. I.; Kell, D. B. *Metabolomics* **2006**, *2*, 171–196.

[5] Brown, M., Dunn, W. B., Ellis, D. I., Goodacre, R., Handl, J., Knowles, J. D., O’Hagan, S., Spasic, I. Kell, D. B. *Metabolomics* **2005**, *1*, 39−51.

[6] Trygg, J.; Holmes, E.; Lundstedt, T. *J. Proteome Res.* **2007**, *6*, 469–479.

[7] Sumner, L. W.; Amberg, A.; Barrett, D.; Beale, M. H.; Beger, R.; Daykin, C. A.; Fan, T. W. M.; Fiehn, O.; Goodacre, R.; Griffin, J. L.; Hankemeier, T.; Hardy, N.; Harnly, J.; Higashi, R.; Kopka, J.; Lane, A. N.; Lindon, J. C.; Marriott, P.; Nicholls, A. W.; Reily, M. D.; Thaden, J. J.; Viant, M. R. *Metabolomics* **2007**, *3*, 211–221.

[8] Dettmer, K.; Aronov, P. A.; Hammock, B. D. *Mass Spectrom. Rev.* **2007**, *26*, 51–78.

[9] Dunn, W. B. *Phys. Biol.* **2008**, *5*, 011001.

[10] Fiehn, O. *TrAC, Trends Anal. Chem.* **2008**, *27*, 261–269.

[11] Lisec, J.; Schauer, N.; Kopka, J.; Willmitzer, L.; Fernie, A. R. *Nat. Protoc.* **2006**, *1*, 387–396.

[12] Bruce, S. J.; Jonsson, P.; Antti, H.; Cloarec, O.; Trygg, J.; Marklund, S. L.; Moritz, T. *Anal. Biochem.* **2008**, *372*, 237–249.

[13] De Vos, R. C. H.; Moco, S.; Lommen, A.; Keurentjes, J. J. B.; Bino, R. J.; Hall, R. D. *Nat. Protoc.* **2007**, *2*, 778–791.

[14] Dunn, W. B.; Broadhurst, D.; Brown, M.; Baker, P. N.; Redman, C. W. G.; Kenny, L. C.; Kell, D. B. *J. Chromatogr., B: Anal. Technol. Biomed. Life Sci.* **2008**, *871*, 288–298.

[15] Metz, T. O.; Zhang, Q. B.; Page, J. S.; Shen, Y. F.; Callister, S. J.; Jacobs, J. M.; Smith, R. D. *Biomarkers Med.* **2007**, *1*, 159–185.

[16] Beckonert, O.; Keun, H. C.; Ebbels, T. M. D.; Bundy, J. G.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. *Nat. Protoc.* **2007**, *2*, 2692–2703.

[17] Dunn, W. B.; Bailey, N. J. C.; Johnson, H. E. *Analyst* **2005**, *130*, 606–625.

[18] Lenz, E. M.; Wilson, I. D. *J. Proteome Res.* **2007**, *6*, 443–458.

10.1021/ac9011599 CCC: $40.75 © 2009 American Chemical Society
Published on Web 07/16/2009

be compiled using several platforms, and multiplatform studies are attracting increasing interest.[19,20]

Following the first direct coupling of GC and MS systems,[21] researchers were quick to apply the technology to biomedical problems.[22] GC/MS is applicable to metabolites which can be volatilized after chemical derivatization, such as fatty acids, organic acids, amino acids, monosaccharides, prostaglandins, steroids, and catecholamines.[23] Metabolite profiling by GC and GC/MS was first studied in the 1960s,[24−26] but its full potential could not be realized until the continuous acquisition of full-scan mass spectra was feasible,[27] most recently using time-of-flight (TOF) mass analyzers at acquisition rates of 10−20 Hz. Such data provide the ability to deconvolve the mass spectra of closely eluting chemical species if the spectra are sufficiently distinct. A significant early application of GC/MS in metabolic profiling was its contribution to the identification of approximately 250 organic acids in urine and characterization of their relationship to organic acidurias and inborn errors of metabolism.[28,29] In the early 21st century, metabolomic research initially focused on plants but rapidly expanded into other areas. Recent metabolomics investigations of clinical interest using GC/MS have included studies of metabolite storage stability in serum and urine,[30] the development of analysis strategies for the plasma metabolome,[31] urinary metabolite profiling,[32] heart failure,[33] pre-eclampsia,[34] diabetes,[35] ovarian cancer using carcinoma tissue,[36] and kidney cancer.[37] However, most of these studies have been performed on comparatively small numbers of samples analyzed together over a short time period.

In order to improve the statistical validity of the acquired data, and to facilitate epidemiological studies, it is highly desirable to employ data set sizes composed of thousands of samples, which requires reproducible analyses over time scales ranging from several months to years. The Human Serum Metabolome project (HUSERMET http://www.husermet.org/), being conducted by the University of Manchester, AstraZeneca, and GlaxoSmithKline, is a study which presents this particular challenge. With employment of a multiplatform approach, this project aims to assemble comprehensive serum metabolic profiles for over 5000 individuals, using samples collected over a 3−5 year period. Raw and processed analytical data, together with clinical and physiological metadata for the subjects, will be a highly useful resource.

Previous optimized methods for metabolomic GC/MS[38] allowed approximately 50 chromatograms per instrument to be acquired in a 24 h period, on a system using a 0.25 mm i.d. capillary column. Application of GC/MS to a study on the scale of HUSERMET requires that variation in process performance must be minimized and methods developed to identify and quantify the variation that remains. Quality control (QC) samples are applied for this purpose, as has been described previously,[39,40] to provide a measurement of process variability.[41]

Careful consideration should be given to factors which could affect the reproducibility of data or produce instrument drift including sample preparation, instrument contamination, detector aging, and data processing. As was recently demonstrated with UPLC−MS data,[40] it cannot be assumed that performance drift will affect all reported metabolites equally or that response changes even occur in the same direction. Step changes in performance both within and between batches can also be observed. Time and order of sample analysis could provide significant sources of variability, potentially obscuring the biological variation which we seek to characterize. Our instrumental analysis has been designed so that the equivalent number of samples can be analyzed by UPLC−MS and GC−TOF-MS in 1 week,[40] this analytical experiment being termed a block. Throughout the article, we will use the term “block” to refer to a set of 120 clinical samples from the study (and associated QC and blank specimens) in a specified run order, and the term “batch” to describe a subset of the block, prepared and analyzed as a single sequence by GC−TOF-MS, and nominally representing 24 h of instrument time. Each block was analyzed as four batches, carried out on consecutive days. This block length represents an informed compromise between the need to use a block sufficiently large to support meaningful statistical analysis, yet small enough to allow reproducibility to be maintained within acceptable parameters. Immediately prior to the instrumental analysis of a block, components with short working lives (syringe, inlet liner, and septum) are replaced, the instrument tuned, and satisfactory chromatographic performance confirmed. No further retuning or maintenance activities are permitted until the entire block has been analyzed.

At present, there are no consensus criteria for assessing reproducibility in metabolomic data sets. For bioanalytical method validation, the (U.S.) Food and Drug Administration

[19] Atherton, H. J.; Bailey, N. J.; Zhang, W.; Taylor, J.; Major, H.; Shockcor, J.; Clarke, K.; Griffin, J. L. *Physiol. Genomics* **2006**, *27*, 178–186.

[20] Buscher, J. M.; Czernik, D.; Ewald, J. C.; Sauer, U.; Zamboni, N. *Anal. Chem.* **2009**, *81*, 2135–2143.

[21] Gohlke, R. S. *Anal. Chem.* **1959**, *31*, 535–541.

[22] Eneroth, P.; Ryhage, R.; Hellstrom, K. *J. Lipid Res.* **1964**, *5*, 245.

[23] Chace, D. H. *Chem. Rev.* **2001**, *101*, 445–477.

[24] Gates, S. C.; Sweeley, C. C. *Clin. Chem.* **1978**, *24*, 1663–1673.

[25] Gates, S. C.; Dendramis, N.; Sweeley, C. C. *Clin. Chem.* **1978**, *24*, 1674–1679.

[26] Tanaka, K.; Budd, M. A.; Efron, M. L.; Isselbac, K. J. *Proc. Natl. Acad. Sci. U.S.A.* **1966**, *56*, 236–242.

[27] Hites, R. A.; Biemann, K. *Anal. Chem.* **1970**, *42*, 855.

[28] Blau, N.; Duran, M.; Blaskovics, M. E. *Physician’s Guide to the Laboratory Diagnosis of Metabolic Diseases*, 1st ed.; Chapman and Hall: London, 1996.

[29] Chalmers, R. A.; Lawson, A. M. *Organic Acids in Man: Analytical Chemistry, Biochemistry and Diagnosis of the Organic Acidurias*; Chapman and Hall: London, 1982.

[30] Dunn, W. B.; Broadhurst, D.; Ellis, D. I.; Brown, M.; Halsall, A.; O’Hagan, S.; Spasic, I.; Tseng, A.; Kell, D. B. *Int. J. Epidemiol.* **2008**, *37*, 23–30.

[31] Jiye, A.; Trygg, J.; Gullberg, J.; Johansson, A. I.; Jonsson, P.; Antti, H.; Marklund, S. L.; Moritz, T. *Anal. Chem.* **2005**, *77*, 8086–8094.

[32] Pasikanti, K. K.; Ho, P. C.; Chan, E. C. Y. *Rapid Commun. Mass Spectrom.* **2008**, *22*, 2984–2992.

[33] Dunn, W.; Broadhurst, D.; Deepak, S.; Buch, M.; McDowell, G.; Spasic, I.; Ellis, D.; Brooks, N.; Kell, D.; Neyses, L. *Metabolomics* **2007**, *3*, 413–426.

[34] Kenny, L. C.; Dunn, W. B.; Ellis, D. I.; Myers, J.; Baker, P. N.; Kell, D. B. *Metabolomics* **2005**, *1*, 227–234.

[35] Major, H. J.; Williams, R.; Wilson, A. J.; Wilson, I. D. *Rapid Commun. Mass Spectrom.* **2006**, *20*, 3295–3302.

[36] Denkert, C.; Budczies, J.; Kind, T.; Weichert, W.; Tablack, P.; Sehouli, J.; Niesporek, S.; Konsgen, D.; Dietel, M.; Fiehn, O. *Cancer Res.* **2006**, *66*, 10795–10804.

[37] Kind, T.; Tolstikov, V.; Fiehn, O.; Weiss, R. H. *Anal. Biochem.* **2007**, *363*, 185–195.

[38] O’Hagan, S.; Dunn, W. B.; Brown, M.; Knowles, J. D.; Kell, D. B. *Anal. Chem.* **2005**, *77*, 290–303.

[39] van der Greef, J.; Martin, S.; Juhasz, P.; Adourian, A.; Plasterer, T.; Verheij, E. R.; McBurney, R. N. *J. Proteome Res.* **2007**, *6*, 1540–1559.

[40] Zelena, E.; Dunn, W. B.; Broadhurst, D.; Francis-McIntyre, S.; Carroll, K. M.; Begley, P.; O’Hagan, S.; Knowles, J. D.; Halsall, A.; HUSERMET Consortium; Wilson, I. D.; Kell, D. B. *Anal. Chem.* **2009**, *81*, 1357–1364.

[41] Sangster, T.; Major, H.; Plumb, R.; Wilson, A. J.; Wilson, I. D. *Analyst* **2006**, *131*, 1075–1078.

(FDA) recommends that for a single analyte test, 60% of the QC standards should lie within 15% of their mean,[42] with a more relaxed criterion (within 20%) applied to analytes at or near their limit of quantitation (LOQ). It can be anticipated that a substantial fraction of the data generated by a metabolomic method will report peaks at or near their LOQ, as a consequence of the aim to produce comprehensive metabolite profiles. Furthermore, the FDA recommendations relate to random variation around a constant mean, while the considerations discussed above lead us to anticipate time-related drift in instrument response. In order to provide a simple basis for comparison, we propose to use a relative standard deviation (RSD) limit of 30% for all QC injections (excluding injection failures) within a block as an acceptance criterion for reporting individual metabolites in the current work, while noting that future users of the HUSERMET data will be able to adjust this criterion to fit the needs of their own research.

In this paper we describe the GC−TOF-MS assay developed for the analysis of serum samples in the HUSERMET project, demonstrate its suitability for metabolomic investigations, and present preliminary data on its stability in a long-term application.

Downloaded by UNIV MANCHESTER on August 14, 2009
Published on July 16, 2009 on http://pubs.acs.org | doi: 10.1021/ac9011599

MATERIALS AND METHODS

This section describes the procedures used for analysis of the HUSERMET study samples. Variations on these procedures for specific validation experiments are described later.

Materials. All materials were purchased from Sigma-Aldrich (Gillingham, U.K.) unless otherwise stated. Pyridine (extra dry), hexane, methoxylamine hydrochloride, and *N*-methyl-*N*-trimethylsilyl-trifluoroacetamide (MSTFA) were obtained from Acros Organics (Loughborough, U.K.).

The internal standards malonic acid-d_2, succinic acid-d_4, and glycine-d_5 were purchased from Sigma-Aldrich (Gillingham, U.K.). Citric acid-d_4, $^{13}C_6$-D-fructose, L-tryptophan-d_5, L-lysine-d_4, L-alanine-d_7, stearic acid-d_{35}, benzoic acid-d_5, and octanoic acid-d_{15} were purchased from Cambridge Isotopes Inc. (Hook, U.K.). Sterile filtered human serum was purchased from Sigma-Aldrich (Gillingham, U.K.) for use as the biological QC material. All of the work reported here was conducted using a single production batch of serum.

Sample Selection and Deproteinization. Serum samples from 5000 healthy individuals are currently being collected over the course of 4 years in the HUSERMET project with the assistance of North West Primary Care Trusts, GlaxoSmithKline, and the European Male Aging Study. The scale of the project demands that sample collection and analysis be performed concurrently. Blocks of 120 samples for analysis are randomly selected from current holdings in our archive, using an automated scheduling system developed by us. This scheduler specifies both the samples and their sequence within the block, and the rationale underlying the block design has been described.[40]

Deproteination and drying of a block of samples is conducted in four batches to conform to the capacity of the vacuum concentrator. Samples selected for the block are randomly assigned to each of the four batches, which also include the preparation of QC serum aliquots. Metabolite extracts for mass spectrometric analysis were prepared by the addition of 200 μL of internal standard solution (0.167 ± 0.009 mg mL^{-1} of each standard), and 1200 μL of methanol to 400 μL of serum, either a study specimen or commercially purchased human serum ("QC Serum") as previously described.[44] Samples were vortexed (15 s) and centrifuged at 15 800*g* for 15 min at room temperature, and the resulting supernatant split between four 2 mL centrifuge tubes (370 μL each, corresponding to 100 μL serum), and dried for 16 h in a vacuum concentrator (HETO VR MAXI with RVT 4104 refrigerated vapor trap; Thermo Life Sciences, Basingstoke, U.K.). Extracts were stored at 4 °C until required for GC−TOF-MS analysis. Saline blanks were prepared using a similar procedure, with 400 μL of serum substituted with 0.7% w/v saline solution. Parallel preparation of four aliquots is convenient for our multiplatform study, but the procedures could readily be adapted to produce single or duplicate aliquots if desirable.

Derivatization of Metabolites for GC−TOF-MS Analysis. For analysis, the scheduled block sequence was reconstructed from the prepared batches and divided into four sequential 24 h analysis batches. QC serum (6 aliquots) and saline blanks (2 aliquots) were derivatized simultaneously. The dried extracts were redissolved in 50 μL of 20 mg mL^{-1} *O*-methoxylamine hydrochloride in pyridine, vortexed, and incubated at 80 °C for 15 min in a dri-block heater. A volume of 50 μL of MSTFA was then added and the extracts incubated at 80 °C for a further 15 min. On completion, 20 μL of retention index marker solution was added (0.3 mg mL^{-1} docosane, nonadecane, decane, dodecane, and pentadecane in pyridine) prior to centrifugation at 15 800*g* for 15 min. The resulting supernatant (90 μL) was transferred to GC/MS vials for analysis.

GC−TOF-MS Analysis. Each 24 h batch of analyses consisted of a five injection "lead-in" or conditioning sequence using QC serum, followed by the samples interspersed with QC and blank specimens. The batch sequence is shown in Supplementary Table 1 in the Supporting Information. The effect of the conditioning injections appears to be minor (Supplementary Table 2 in the Supporting Information). Analyses were carried out using a Leco Pegasus III (4D) GC × GC/MS in GC/MS mode (Leco Corp., St. Joseph, MO), with a Gerstel MPS-2 autosampler (Gerstel, Baltimore, MD) and an Agilent 6890N gas chromatograph with a split/splitless injector and Agilent LPD split-mode inlet linear (Agilent Technologies, Stockport, U.K.). The previously published operating parameters[38] were modified due to hardware differences between the two instruments. Oven cooling was significantly slower due to components required for GC × GC operation, and to maintain sample throughput, the carrier gas flow rate was increased and the initial and final hold times reduced. Injection parameters were also modified to reflect differences in the injection speed of the current (Gerstel) and previous (Agilent) autosamplers. A 30 m × 0.25 mm × 0.25 μm VF17-MS bonded phase capillary column (Varian, Oxford, U.K.) was used at a constant helium carrier gas flow of 1 mL min^{-1}. Temperature program: 4 min hold at 70 °C, 20 °C/min to 300 °C, 4 min hold. Sample injections (1 μL) were made using an empty hot needle[43]

[42] Centre for Drug Evaluation and Research (CDER). FDA Guidance for Industry, Bioanalytical Method Validation, May **2001**.

[43] Grob, K.; Neukom, H. P. *J. High Resol. Chromatogr. Chromatogr. Commun.* **1979**, *2*, 15–21.

[44] Tukey, J. W. *Ann. Math. Stat.* **1962**, *33*, 1–67.

technique. Other autosampler parameters were chosen to minimize problems with the entrainment of bubbles into the syringe. The injector was operated at 280 °C with a 4:1 split ratio, and a 25 mL min^{-1} gas saver flow switched on after 30s. The transfer line was maintained at 240 °C. The mass spectrometer was operated at 70 eV ionization energy with a source temperature of 220 °C, acquiring m/z 45−600 at 20 Hz.

Raw Data Processing. The process used was based on the "Compare" capability of LECO's ChromaTOF v3.25 software (Leco Corp., St. Joseph, MO). A set of reference spectra are compiled for a list of serum metabolites, with all subsequent samples searched against the reference table generated. Where possible, peak identities were assigned on the basis of mass spectral similarity to NIST library entries or mass spectral and retention index comparison with an in-house metabolite library generated from authentic standards at the University of Manchester.

In order to produce an unbiased set of targets, all peaks present in a representative QC sample which met specified criteria for signal/noise (S/N) ratio and chromatographic peak width were considered for inclusion in the reference table. The list was manually edited to ensure the mass spectra conformed to the expected fragmentation patterns for TMS derivatives and to remove duplicates or deconvolution artifacts. Appropriate peak detection parameters in ChromaTOF are strongly dependent on the chromatographic performance and sensitivity of the instrument, and a peak width of 1.8 s and S/N ratio of 100:1 proved suitable. The reference table comprised 200 peaks (metabolites) obtained from typical QC serum chromatograms, an additional 55 peaks found in one or more of 24 chromatograms from representative clinical samples, and a further 13 entries were added comprising retention index markers and internal standards.

Once the reference table had been generated, supporting information for each peak was specified, e.g., quantitation ions and acceptable tolerances for mass spectral similarity and retention indices. Internal standards were used as the basis for relative quantitation: for each metabolite, an internal standard was assigned within the reference table, allowing ion ratio data to be reported automatically. The use of retention markers allowed retention indices to be calculated for all other peaks in the chromatogram by interpolation. For each block of 120 clinical samples, a new retention index table was manually created, allowing significant changes in absolute retention time to be compensated.

Statistical Analysis. Both univariate and multivariate analysis was performed on the ion ratio data sets. Univariate procedures were performed using Microsoft Excel 2003, and multivariate data analysis was performed using Matlab R2008.a (MathWorks, Inc., MA). For the multivariate analysis, all peaks with more than 20% missing values were removed from the analysis. Outliers were suppressed using 95% winsorisation.[44] The remaining few missing values were replaced with median values. All metabolite ion ratio data were normalized to unit variance and mean centered before further analysis. Unsupervised multivariate analysis was performed using principal components analysis (PCA),[45] and supervised multivariate analysis was performed using principal components followed by canonical variates analysis (PC-CVA).[46] The PC-CVA models that were constructed were cross-validated by randomly removing one-third of the available data from the training set and using it as a test set, where these data were first projected into PCA scores space and then into CVA scores space from the models constructed on the training data.

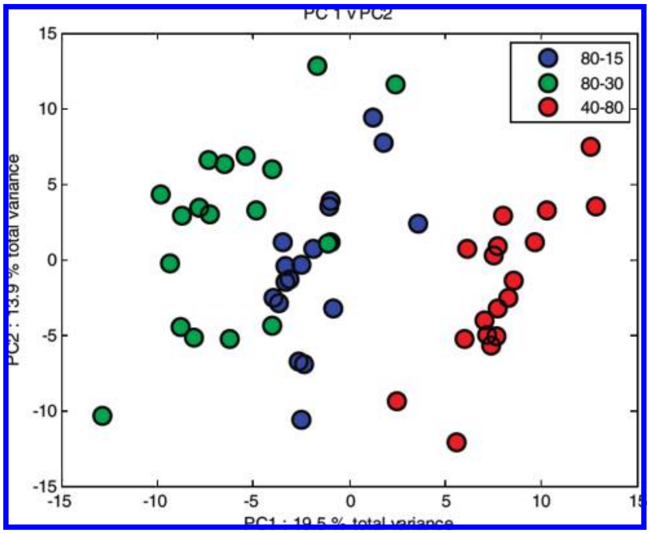

Figure 1. PCA scores biplot of three sets of derivatization reaction conditions: for each stage (oxime formation and trimethylsilylation), red, 80 min at 40 °C; blue, 15 min at 80 °C; and green, 30 min at 80 °C.

RESULTS AND DISCUSSION

Effect of Derivatization Reaction Conditions. A useful discussion of the ease of formation and subsequent stability of oxime/TMS derivatives of various metabolite classes has recently been published,[47] for a reagent system comparable to that used in this work. The selection of derivatization conditions is intended to lead to reproducible conversion of the targets to their most highly derivatized states while minimizing the extent of side reactions and product degradation, giving a simple stable relationship between the signal measured and the amount of that metabolite in the original sample. Complete derivatization of all metabolites of interest is usually achievable in the targeted analysis of chemically similar compounds but is challenging in a metabolomic analysis which is intended to be comprehensive and where the identities of some of the metabolites reported remain unknown.

To assess the impact of derivatization conditions, batches of 18 QC serum aliquots were prepared. Three sets of reaction conditions were tested: for each reaction step 80 min at 40 °C (as used in previous work[30]), 30 min at 80 °C, and 15 min at 80 °C, with all samples analyzed as a single randomized block. Considering all features reported (250), median relative standard deviations (RSDs) were 18.4%, 18.3%, and 37.5%, for 15 min/80 °C, 80 min/40 °C, and 30 min/80 °C respectively, indicating lower repeatability for the 30 min/80 °C conditions. In Figure 1, where the data have been subjected to PCA, this is reflected in the greater dispersion of the 30 min/80 °C data points along the PC1 axis. The three

(45) Jolliffe, I. T. *Principal Components Analysis*; Springer-Verlag: New York, 1986.

(46) Krzanowski, W. J. *Principles of Multivariate Analysis: A User's Perspective*; Oxford University Press: Oxford, U.K., 1988.

(47) Koek, M. M.; Muilwijk, B.; van der Werf, M. J.; Hankemeier, T. *Anal. Chem.* **2006**, *78*, 1272–1281.

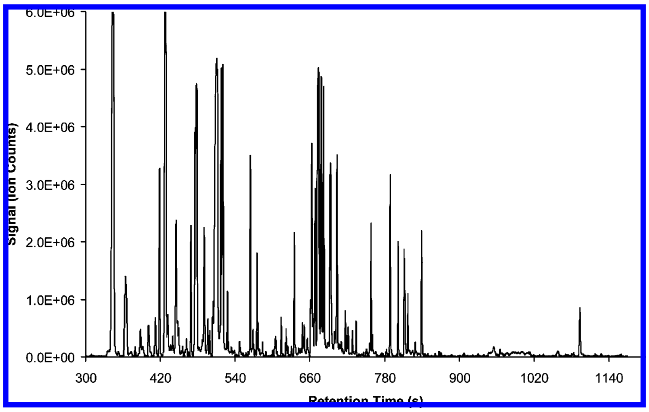

Figure 2. Total ion current chromatogram (baseline corrected within ChromaTOF using the default parameters) for a typical serum sample using the conditions described.

reaction condition sets are displaced from each other and distinguishable along PC1 but show similar variability along PC2. In the loadings plot for PC1 (data not shown), peaks with high positive loadings included glutamine, glutamic acid, and pyroglutamic acid, which are known to have the potential to interconvert during extraction and derivatization[48,49] and might therefore be particularly responsive to changes in reaction conditions. Inspection of the data for known metabolites suggested that 15 min/80 °C and 80 min/40 °C corresponded to a similar overall extent of reaction, as similar responses were observed. It was anticipated that 30 min/80 °C would result in reaction closer to completion for metabolites requiring more "forcing" reaction conditions. Although the data (Supplementary Table 3 in the Supporting Information) suggest this may have occurred, interpretation is confounded by the reduced repeatability, which requires further study. The fast 15 min/80 °C method is convenient for batch processing of samples and was adopted for the remainder of the work reported here, and a typical total ion chromatogram is shown in Figure 2.

Stability of Derivatized Extracts. Five sets of QC serum samples were derivatized (15 min at 80 °C) and held in sealed GC vials for varying times prior to analysis. The sets were combined as a single randomized batch for analysis. Set 1, 2, 3, 4, and 5, respectively, are samples stored for varying times. Analysis of set 1 commenced within 4 h of derivatization completion. Sets 2, 3, 4, and 5 were analyzed 24, 48, 72, and 120 h after completion of derivatization, respectively.

No clear separation of the data was observed using PCA in the first four PCs which accounted for 60% of the total variance (data not shown). With the use of supervised PC-CVA (Figure 3), differences could be observed between all sample classes, supporting our standard protocol which stipulates that the batch should be commenced within 4 h of derivatization completion, allowing batch completion within 30 h. These data show a relatively small separation on CV1 for sets 1−3 compared to sets 4 and 5.

Reference Table Quality. Within the method applied for reference table generation (LECO ChromaTOF v3.25) and data reduction, user-adjustable parameters including mass spectral match, retention index window, and quantification ion influence the reliability of subsequent peak reporting. Each sample chromatogram is searched for all peaks in the reference table after the initial deconvolution process. The compilation of the reference table is described in the Materials and Methods.

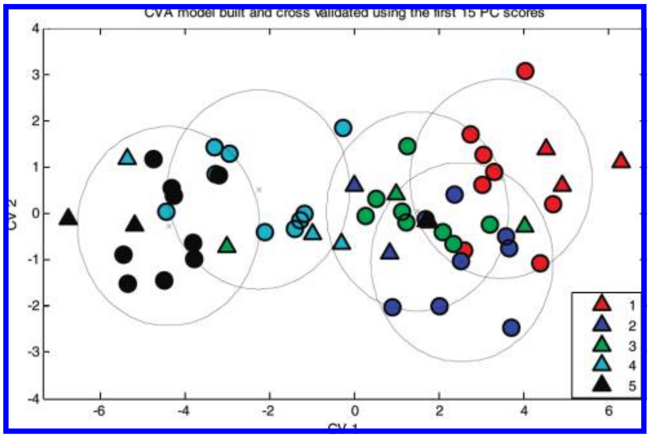

Figure 3. Stability on storage of derivatized samples using a trained CVA model: effect of 4 (red), 24 (blue), 48 (green), 72 (light blue), or 120 (black) hour delay prior to starting batch analysis. filled circles, training set; filled triangles, test set; large circles, 95% confidence regions.

To help assess which peaks relate to metabolites that can meaningfully be measured in the samples, a differential serum volume experiment was performed. A series of samples was prepared using QC serum, in which all quantities except the serum volume were constant. The serum volume was varied in five steps from 40 to 180 μL, which represent 40−180% of the sample quantity used as standard. Of 200 chromatographic peaks initially selected from the QC serum, 130 showed a scaling response to serum volume (r^2 value greater than 0.75), while 70 did not. A number of the scaling peaks were observed to maintain a linear relationship over the range 40−140 μL but to depart from this at the 180 μL level, suggesting that 100 μL serum extracts were appropriate for the main study. Of the 70 peaks which failed to scale, 43 had significant contributions from background in the blank saline samples. Across the nine blocks of analyses discussed below, peaks which passed the scaling test had a median response in the blank of 1.6% their response in the QC serum, while for the group which failed to scale with serum volume, median response in the blank was 26.5% (see Supplementary Figure 2 in the Supporting Information). Other peaks which failed the scaling test included some that saturated the detector (e.g., glucose), formed unstable derivatization products (e.g., urea), or were present at concentrations close to their limits of detection and may have failed to establish a scaling response over the relatively narrow range of serum volumes used (limited by column overloading of major components). Full-scan acquisitions are intrinsically less selective than the selected ion monitoring commonly used for targeted assays and more susceptible to background contributions. It should be emphasized that a "serum scaling failure" may be a metabolite of biological interest which could be measured if contributions to the background could be characterized and reduced. As a group, the peaks which failed

[48] Darmaun, D.; Manary, M. J.; Matthews, D. E. *Anal. Biochem.* **1985**, *147*, 92–102.

[49] Gehrke, C. W.; Nakamoto, H.; Zumwalt, R. W. *J. Chromatogr.* **1969**, *45*, 24.

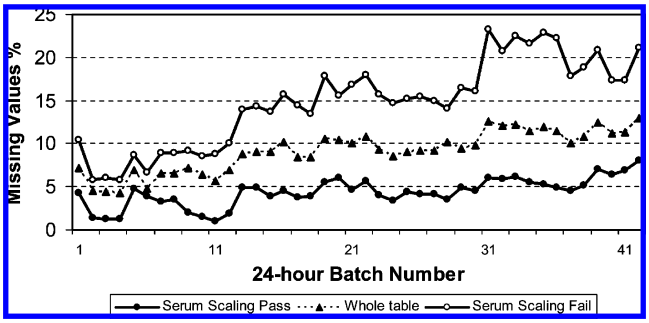

Figure 4. Variation in missing values over 42 consecutive 24 h batches. The peak features which failed the serum scaling test (see text for details) accumulated a higher and increasing proportion of missing values than did those that passed the scaling test.

Table 1. Median Relative Standard Deviation (%) for Peaks Measured in 575 QC Serum Injections November 17, 2007 to April 17, 2008

	raw area	ratio vs octanoic acid-d_{15}	ratio vs best internal standard
whole reference table (255 peaks)[a]	51.1	45.3	43.0
"serum scaling passes" only (130 peaks)	39.6	35.6	33.3

[a] The whole reference table includes 130 "serum scaling passes", 70 "serum scaling fails", and 55 peaks not tested in the scaling experiment.

the scaling test accumulated a higher rate of missing values than those that scaled with serum volume, as successive blocks of the main study samples were analyzed (Figure 4). This is again consistent with a variable contribution to their measured responses from background over the 5 month period when the blocks were analyzed. The scaling experiment provided an early indication of data quality which could otherwise only be obtained by assessing the method's performance over an extended period.

With the use of a biological sample to generate the reference table, the mass spectra extracted for low-concentration metabolites may include m/z peaks derived from partially coeluting species and background. Furthermore, most TMS derivatives generate high intensity fragments at m/z 73, 75, and 147, which are entirely related to the trimethylsilyl group, while ions characteristic of the metabolite can be present at low relative abundance. As a result, mass spectral similarity can be high for structurally distinct metabolites and the deconvolution algorithm is likely to be challenged by coeluting peaks where a significant fraction of the mass spectrum is similar for both coeluants. This was apparent with the isotopically labeled standards and their analogues (e.g., alanine and alanine-d_7), for which automatic peak finding was initially unreliable. Manually editing the reference spectra to remove noncharacteristic ions was found to be effective and could be an appropriate approach with other problematic peaks. An alternative, based on explicitly searching for only one or two m/z signals known to be characteristic of each target metabolite, has recently been described,[15] while a range of other tools are being offered,[50,51] developed, and studied.[10,52] In the current context, where our focus is primarily on assessing procedural robustness, ChromaTOF was judged to provide a reliable option for data reduction.

As will be demonstrated below, the current version of the reference table allows biologically meaningful data to be derived from raw chromatograms. However, the observations described suggest that a more sophisticated approach to reference table generation could lead to improvement. It is possible that by removing noncharacteristic TMS ions (m/z 73, 75, and 147) from consideration, the deconvolution process would be more robust where coelution occurs. A more challenging approach would be to attempt to generate a "consensus" reference table from biologically identical samples (e.g., the QC serum) run at different times through the study and in this way identify and eliminate peaks derived from background. These approaches could usefully be followed for future study.

Internal Standard Selection. Inclusion of a known quantity of an internal standard in every sample allows a series of analyses to be corrected for variation in the volume injected. Further, the vaporization process within split and splitless injectors[53] results in variable transfer efficiency,[54] both for components of different volatilities in a single sample and for the same component in different samples if the evaporation profile varies from injection to injection. Use of a set of internal standards of differing volatilities and retention behavior allows compensation for variation in transfer efficiency. When chemical derivatization is employed, a set of internal standards of differing chemical reactivities can allow compensation for variations in reaction conditions. The internal standards used here were chosen to reflect the anticipated range of functional groups, degree of polyfunctionality, and volatility.

For the purpose of relative quantitation, an internal standard was selected from our 11 candidates for every peak present in the reference table by selecting the candidate which minimized the relative standard deviation for that peak, using 575 QC serum injections in study blocks run over a period of 5 months. Of the candidate standards considered, only succinic acid-d_4, citric acid-d_4, $^{13}C_6D$-fructose, L-tryptophan-d_5, L-alanine-d_7, benzoic acid-d_5, and octanoic acid-d_{15} were selected for long-term use by this criterion. As Table 1 demonstrates, the use of a single internal standard (e.g., octanoic acid-d_{15}) provides an improvement in reproducibility. Initially octanoic acid-d_{15} was used as a global internal standard, and the data presented here were recalculated using the optimal internal standards for each metabolite once these were determined for each individual metabolite feature.

Demonstration of Ability to Detect Metabolic Differences in One Block. A sample set was constructed from 60 study samples. Aliquots (50 μL) from each sample were pooled, and 60 pooled serum aliquots were prepared. Of these, 12 samples ("spike 1") were spiked with glutaric acid, citric acid, alanine, glycine, leucine, phenylalanine, and tryptophan at a concentration of 0.16 mg mL^{-1} per component. An additional 12 QC samples ("spike

[50] http://chemdata.nist.gov/mass-spc/amdis.
[51] http://masspec.scripps.edu/xcms/xcms.php.
[52] Jonsson, P.; Gullberg, J.; Nordstrom, A.; Kusano, M.; Kowalczyk, M.; Sjostrom, M.; Moritz, T. *Anal. Chem.* **2004**, *76*, 1738–1745.
[53] Grob, K. *Split and Splitless Injection in Gas Chromatography*, 3rd ed.; Huthig: Heidelberg, Germany, 1993.
[54] Kaufman, A. E.; Polymeropoulos, C. E. *J. Chromatogr.* **1988**, *454*, 23–36.

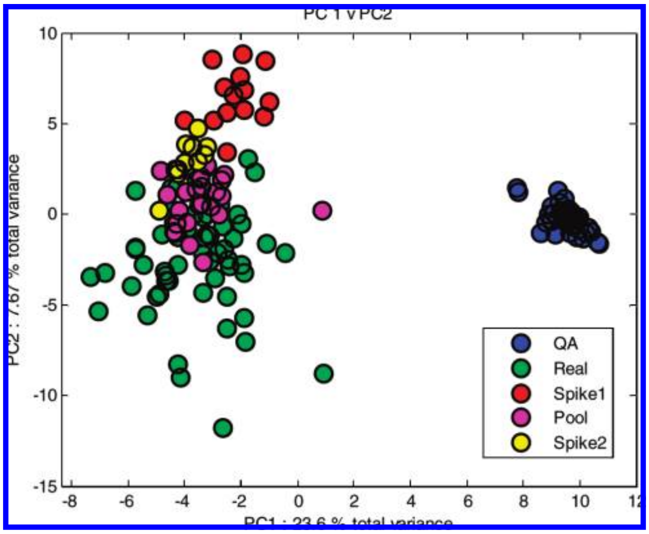

Figure 5. Demonstration of response to biological variability. PCA scores plot showing discrimination of QC serum (blue), pooled serum (pink), individual study specimens (green), and two spiked pooled serum samples: spike 1 (red) and spike 2 (yellow) as described in the main text.

2") were spiked with caffeine and nicotine each at a concentration 0.16 mg mL^{-1}. The remaining aliquots were analyzed unspiked. All of the components in spike 1 were known to be present in the reference table, and so it was expected that it would be readily discriminated from the pooled serum. In contrast, spike 2 included only one component (caffeine) in the table (nicotine was not present as a target in the reference list), providing a more stringent test. These samples were complemented with 45 (Sigma) QC serum samples and blanks to replicate the block structure used for the main clinical study. The study samples were randomized across the block, while the pooled and spiked samples were uniformly distributed.

Peaks which had missing value rates >20% were removed from the data set prior to PCA. A total of 145 peaks remained from the original 255 in the reference table. The PC1 vs PC2 plot (Figure 5) shows the QC serum is well separated from the experimental samples on PC1, with a tighter grouping than either the pooled serum or the study samples. The QC serum is subjected to a sterile filtration process which is not performed on our study samples, and its clear discrimination from them may reflect changes (components either removed or added) during this process. The clinical, pooled, and spike 1 samples are separated primarily along PC2. Although spike 2 is not fully separated from either the clinical samples or the pooled samples, this was anticipated as spike 2 was specifically intended to provide a severe test. The clinical samples show a greater dispersion across the PC1−PC2 plane than either spike 1 or the pooled serum samples. Both spike 1 and the pooled serum were less closely grouped than the QC serum, which may simply reflect the difficulty of producing uniformly pooled serum in small volumes. When the loadings plot is considered (data not shown), the six table entries given the highest positive loading on PC2 are identified as tryptophan (two peaks), glycine (two peaks), phenylalanine, and hydroxypyridine. In respect to the amino acids, this confirms that the multivariate analysis is responding to chemical variation built into our test. Hydroxypyridine is, however, a surprising candidate to provide discriminating power in this experiment and its high loading may indicate imprecision or crosstalk in the reference table.

Assessment of Reproducibility for Nine Blocks of Samples. The results described above have allowed us to construct an appropriate standard operating procedure (SOP) and demonstrate that a block composed of four separate batches analyzed over a 1 week period can be treated as a single experiment. For large scale studies in which experiments are conducted over a number of months as a series of blocks, stable chromatographic performance is essential. Data from 5 months of analyses was assessed for chromatographic stability and reproducibility. The median peak width for the sample blocks 1−8, equivalent to 1 080 serum samples and a minimum of 1 600 injections, increased from 1.20 s (block 1) to 1.25 s (block 8). This demonstrates very little loss in chromatographic resolution. A new column was installed prior to block 9. Overall, the retention times for the retention index markers varied by ±1.5% over the nine blocks (see Supplementary Figure 1 in the Supporting Information). However, as the figure demonstrates, the major retention time change in the data does

Table 2. Reproducibility of Peak Response Ratio Data for QC Serum for a Single Batch and for Nine Blocks

	median RSD (%) of internal standard ratio		total no. of peaks (metabolites) with RSD (%) better than		
	whole reference table (255 peaks)	serum scaling passes (130 peaks)	20%	25%	30%
24 h batch (average of all batches in blocks 1−9)	23.0	18.4	107	133	152
4 day block (60[a] QC), average	29.2	22.5	73	104	129
block 1	28.1	22.1	73	106	135
block 2	38.2	33.7	26	38	76
block 3	31.3	24.7	67	91	117
block 4	30.0	22.3	91	108	125
block 5	27.9	23.2	65	110	135
block 6	28.5	22.2	79	102	130
block 7	27.8	22.5	68	111	130
block 8	32.5	21.6	85	104	120
block 9	29.2	22.6	75	99	129
all QC in blocks 1−9	43.0	33.3	17	49	72

[a] Nominal number of replicates specified by the block design; the true number used varies slightly from block to block, reflecting reinjections and failed injections. The data reported includes a scheduled column replacement between blocks 8 and 9.

not correspond to column substitution. Variability between column batches appears small compared with variation introduced by factors such as injector maintenance. With the use of an updated retention index table for each block, retention indices were highly reproducible, median RSDs were typically 0.05% for the "scaling pass" group and 0.10% for all features reported within a block. Within-block reproducibility and peak detection consistency were also assessed using the data acquired for QC samples present in each block of samples and are summarized in Table 2. With the exception of block 2, consistent within-block performance is observed throughout, with similar median RSD for response ratios, and numbers of detected metabolite features (peaks) achieving a given level of reproducibility, for each block. Reference to the instrument log reveals problems with injection reproducibility due to entrained bubbles during block 2, not encountered in other blocks. This emphasizes the need to reinforce the prescriptive aspects of an SOP with postacquisition assessment of the data sets against acceptance criteria. Those peaks which were previously identified as scaling with serum volume demonstrated better reproducibility than the reference table as a whole. This is consistent with the hypothesis that for those peaks which did not scale with serum volume, a significant and varying fraction of the observed signal is not derived from the sample under study but from other sources.

CONCLUSIONS

The procedure described has been demonstrated to generate reproducible data at the block level and consistent long-term performance over a number of blocks. The inclusion of frequent analysis of replicate and biologically identical QC specimens provides both data validation capabilities through a quantification of variability and a foundation for interpolation-based approaches to alignment of blocks to form a single data set on a metabolite-by-metabolite basis.

The choices of sample volume, solvent volume, drying conditions, and subsequent derivatization conditions have proved highly robust, with derivatization failures occurring at fewer than one per block, which has beneficial consequences for the overall stability of chromatographic performance. The useful storage time available after derivatization (30 h) is a reflection of the large excess of derivatizing agent (MSTFA), which protects the derivatized samples from hydrolysis. We have achieved working lives (to the above standards) in excess of 2 000 injections using VF17-MS columns. The mass spectrometer required minimal intervention. The detector had a working life of 12−15 months, and over this interval the voltage needed to achieve the required S/N increased progressively and was correctly adjusted automatically during the tuning procedure run prior to each block. Filaments were found to have working lives in the range 8−12 months. Future efforts to improve the reproducibility of hardware operation could usefully focus on the injection technique, as we have encountered a rate of injection failures of 5−6% and variable injection volumes.

By frequent QC replicates built into the block design, the reanalysis and reprocessing of the data have been facilitated. Provided a metabolite is present at measurable levels in the QC serum, it will be possible to recover analytical reproducibility data retrospectively, even for metabolites not in our current target list. As our comments on reference table quality suggest, the data reduction and reporting method used here might best be regarded as a first draft, since there appears to be scope to improve its reliability either within the proprietary software framework used here or by using alternative approaches. The block length used is compatible with our parallel UPLC−MS investigation.[40] In the context of GC−TOF-MS, while instrument drift does not limit the block length, episodic failure of the inlet septum or the syringe becomes more frequent if blocks requiring longer periods of operation are employed.

The number of peaks reported is low by comparison with other published metabolomic GC−TOF-MS work, 130 at our proposed level of acceptable reproducibility. This reflects a number of factors. Some authors have quoted the initial number of peaks selected for reporting, without providing information on their reproducibility. We have observed that as the size of the data set increases, metrics of reproducibility degrade, as would be expected if the data carry progressive trends and/or step changes in performance. This implies that experiments on a short time scale will produce more comprehensive data sets at any level of acceptable reproducibility. Finally, the logistics of our long-term study have necessitated the use of commercial serum as the QC material. In shorter studies, a sample pooled[41] from all of the study specimens would allow a more representative and comprehensive target list to be compiled. While it is desirable to maximize the number of peaks reported, and by implication the coverage of the metabolome, reporting artifact peaks and relying upon subsequent data analysis to filter these out is not a penalty-free option, as the target compound analysis software may ascribe fractions of the ion current for a metabolite peak to adjacent artifacts. More generally, the commonly used multivariate methods of data visualization (e.g., PCA) will respond to all sources of variability in the data, and so it is highly desirable to minimize the contributions from sources other than the biological system under study.

Comparing our experience with UPLC−MS[40] and GC−TOF-MS, we have identified a different set of factors which influence reproducibility, largely reflecting instrumental design. With both techniques, we have arrived at a preferred block length of 180 injections, but in the case of UPLC−MS this is a consequence of column and source contamination, while in GC−TOF-MS it primarily reflects the probability of mechanical failure by the inlet septum or syringe. In GC−TOF-MS, both the column and source are achieving working lives at least an order of magnitude longer than those achieved by our UPLC−MS method. This reflects lower sample loading, the protection of column and detector from low volatility material which a vaporizing injection technique provides, and the greater tolerance of source contamination by electron impact ionization in comparison with "softer" electrospray ionization. Conversely, the flash vaporization and resultant pressure pulse in the split/splitless inlet employed in GC−TOF-MS present a source of variation (variable transfer efficiency) that is entirely absent in UPLC−MS. This and process variability during chemical derivatization appear to be the factors that reduce the short-term repeatability of a GC−TOF-MS method below than that achievable on UPLC−MS.

In conclusion, we have developed a robust SOP for the collection of GC−TOF-MS data from serum samples and found that the long-term reproducibility of the metabolite data generated was excellent. We would advocate the use of this SOP with suitable QC samples for the assessment of long-term untargeted metabolomic studies.

ACKNOWLEDGMENT

HUSERMET is funded by the U.K. BBSRC, MRC, GlaxoSmithKline, and AstraZeneca. We also thank the BBSRC and EPSRC for financial support of The Manchester Centre for Integrative Systems Biology. P.B., S.F.-M., W.B.D., and D.I.B. contributed equally to the research. Sjaak de Koening, Ralf Löscher, and Lorraine Kay (LECO Europe) provided much useful advice on the use of ChromaTOF software.

SUPPORTING INFORMATION AVAILABLE

Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review May 28, 2009. Accepted June 24, 2009.

AC9011599